STARC-9: A Large-scale Dataset for Multi-Class Tissue Classification for CRC Histopathology

Barathi Subramanian[1,*] **Rathinaraja Jeyaraj**[1,*] **Mitchell Nevin Peterson**[2]
Terry Guo[1] **Nigam Shah**[3] **Curtis Langlotz**[4] **Andrew Y. Ng**[5,6] **Jeanne Shen**[1]
{[1]Department of Pathology, [2]Department of Electrical Engineering, [3]Department of Medicine,
[4]Department of Radiology, [5]Department of Computer Science}, Stanford University, USA
[6]DeepLearning.AI, USA

Abstract

Multi-class tissue-type classification of colorectal cancer (CRC) histopathologic images is a significant step in the development of downstream machine learning models for diagnosis and treatment planning. However, publicly available CRC datasets used to build tissue classifiers often suffer from insufficient morphologic diversity, class imbalance, and low-quality image tiles, limiting downstream model performance and generalizability. To address this research gap, we introduce STARC-9 (STAnford coloRectal Cancer), a large-scale dataset for multi-class tissue classification. STARC-9 comprises 630,000 histopathologic image tiles uniformly sampled across nine clinically relevant tissue classes (each represented by 70,000 tiles), systematically extracted from hematoxylin & eosin-stained whole-slide images (WSI) from 200 CRC patients at the Stanford University School of Medicine. To construct STARC-9, we propose a novel framework, DeepCluster++, consisting of two primary steps to ensure diversity within each tissue class, followed by pathologist verification. First, an encoder from an autoencoder trained specifically on histopathologic images is used to extract feature vectors from all tiles within a given input WSI. Next, K-means clustering groups morphologically similar tiles, followed by an equal-frequency binning method to sample diverse patterns within each tissue class. Finally, the selected tiles are verified by expert gastrointestinal pathologists to ensure classification accuracy. This semi-automated approach significantly reduces the manual effort required for dataset curation while producing high-quality training examples. To validate the utility of STARC-9, we benchmarked baseline convolutional neural networks, transformers, and pathology-specific foundation models on downstream multi-class CRC tissue classification and segmentation tasks when trained on STARC-9 versus publicly available datasets, demonstrating superior generalizability of models trained on STARC-9. Although we demonstrate the utility of DeepCluster++ on CRC as a pilot use-case, it is a flexible framework that can be used for constructing high-quality datasets from large WSI repositories across a wide range of cancer and non-cancer applications. `https://huggingface.co/datasets/Path2AI/STARC-9/tree/main`
`https://github.com/Path2AI/STARC-9/`

1 Introduction

Colorectal cancer (CRC) is the 3rd most common cancer and the 2nd leading cause of cancer-related death worldwide [27]. Histologic evaluation of CRC is essential for diagnosis, prognostication, and therapeutic decision-making. With the growing adoption of digital pathology, computational approaches, particularly those leveraging deep learning, will play an increasingly important role in

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

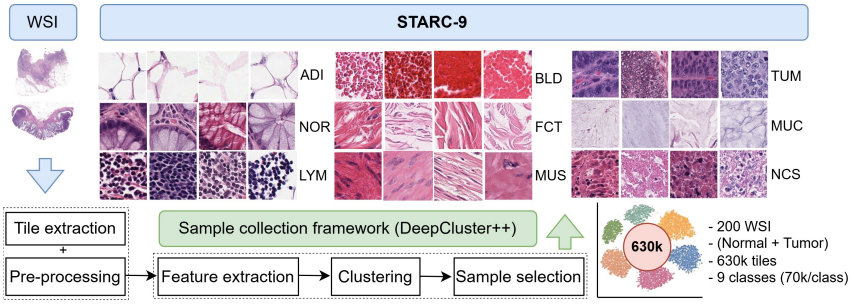


Figure 1: Overview of STARC-9 large-scale dataset generation.

automating and augmenting pathology workflows. Deep learning-based multi-class tissue classification represents one such foundational task in pathology [17], [30], enabling models to distinguish between diverse tissue types such as tumor, normal epithelium, muscle, and necrotic regions, and supporting downstream applications such as tissue segmentation [4], tissue composition analysis [13], biomarker status prediction [46], and survival analysis [15]. Providing pathologists with visually intuitive tissue maps also reduces the mental burden of diagnosis, while enhancing the interpretability of AI-driven insights. However, publicly available CRC datasets for machine learning are limited. The NCT-CRC-HE-100K dataset from Kather et al. [15] represented a significant contribution for multi-class CRC tissue classification. Recently, another dataset (HMU-GC-HE-30K) [24] was made public with different tissue types for developing tissue classifiers. Although this dataset contains images obtained from gastric cancer specimens, many of the tissue classes overlap with those found in CRC. Additional publicly available CRC datasets include the TCGA COAD and READ (The Cancer Genome Atlas - Colorectal Adenocarcinoma and Rectal Adenocarcinoma) [40] whole-slide image (WSI)-level datasets. Despite representing important contributions to the field, these and other currently available histopathologic datasets have been insufficient for building robust, generalizable models for tissue-type classification and other downstream applications for several reasons, including: (1) lack of morphologic diversity, with insufficient representation of the broad range of appearances of different tissue classes [39], (2) class imbalance, where samples from dominant tissue types (e.g., tumor epithelium) far outnumber other clinically significant classes (e.g., mucin or necrosis) [11], and (3) inclusion of non-representative (incorrectly classified) tiles and low-quality (artifact-containing) tiles [18], which hinder model interpretability and degrade downstream task-specific performance. Furthermore, the construction of new pathology datasets is labor intensive, with no standardized framework for capturing sufficient tissue diversity and morphologic variation. All of these represent significant barriers to the development of robust, generalizable tissue classification models.

To address these limitations, we introduce STARC-9 (STAnford coloRectal Cancer), a large-scale, high-quality dataset specifically designed for multi-class tissue classification for CRC histopathology, as well as a novel framework, DeepCluster++, used to construct STARC-9, which can readily be applied to other types of histopathologic WSI. STARC-9 comprises 630,000 non-overlapping tiles (256x256 pixels) systematically extracted using DeepCluster++ from hematoxylin & eosin (H&E)-stained WSI from approximately 200 CRC patients who underwent surgical resection of their CRC at the Stanford University School of Medicine. The following nine clinically relevant tissue classes are uniformly represented in the dataset: adipose tissue (ADI), lymphoid tissue (LYM), muscle (MUS), fibroconnective tissue (FCT), mucin (MUC), necrosis (NCS), blood (BLD), tumor (TUM), and normal mucosa (NOR), with each class containing 70,000 tiles. The overall DeepCluster++ workflow for constructing STARC-9 is illustrated in Figure 1. Initially, tiles were extracted from the WSI and preprocessed to remove background and artifact tiles. Then, feature vectors for the remaining informative tiles were extracted using a CRC-specific autoencoder, pretrained on 100,000 histopathology images. K- means clustering was then employed to group the tiles based on morphologic similarity, wherein each cluster might represent a particular tissue type. To avoid oversampling in dense centroid regions, we partitioned each cluster into equal-frequency distance bins. Sampling across these bins ensured balanced intra-cluster diversity for robust classifier training. Repeating this pipeline across all WSI yielded 70,000 tiles per tissue type, resulting in 630,000 total high-quality tiles. Experienced pathologists then reviewed these samples to verify tissue-type classification accuracy, resulting in a robust, clinically relevant dataset.

The DeepCluster++ framework for dataset construction significantly reduces the time and effort required for tile selection, compared to manual annotation using open-source tools such as QuPath [2].

Traditional region-based sampling often leads to limited morphologic diversity, as pathologists tend to focus on visually similar regions within WSI, resulting in class imbalance and reduced generalizability of downstream models. Manual annotation is also subjective and inconsistent, relying heavily on the pathologist's impression of whether a new region is sufficiently different from a previously annotated region to warrant inclusion in the dataset, making it difficult to ensure comprehensive representation of the entire morphologic spectrum within a WSI. In contrast, DeepCluster++ employs unsupervised clustering to group structurally similar tiles into coherent clusters, regardless of their location within a WSI. As a result, each cluster contains tissue tiles of a similar appearance sampled from diverse regions within the WSI. Sampling tiles from clusters corresponding to the same tissue type in this way enhances intra-class heterogeneity and tissue morphologic diversity (for instance, NOR, TUM, NCS in Figure 1). This enhances dataset quality and increases the generalizability of models trained on the dataset by exposing them to a broad range of tissue morphologies important for downstream clinical applications. Although minimal manual review is still required, this method streamlines the overall dataset collection process, producing high-quality training examples for robust model development. Furthermore, DeepCluster++ is a flexible framework for constructing high-quality datasets which can be applied to both cancer and non-cancer WSI.

In a comprehensive evaluation, we trained both baseline and advanced multi-class classification models on the STARC-9, HMU-GC-HE-30K [24], and NCT-CRC-HE-100K [15] datasets and evaluated their performance on independent Stanford and TCGA-CRC datasets using standard evaluation metrics, including precision, recall, F1-score, and accuracy. The baseline models included ResNet-50, EfficientNet-B7, KimiaNet, and ViT-base, as well as state-of-the-art (SOTA) transformer models such as DeiT-B, Swin Transformer-base, and ConvNeXT-Base. Pathology-specific foundation models, including CTransPath, HiPT, Prov-Gigapath, Path-DINO, CONCH, UNI, Virchow, and VIM4Path, were also benchmarked to evaluate their generalizability after fine-tuning on our dataset. In addition, a custom convolutional neural network (CNN) and a Histo-ViT model (DeiT-B) trained from scratch were also included in the analysis. In addition to these quantitative evaluations, we validated the practical utility of STARC-9 for tile-level segmentation on an independent TCGA-CRC dataset. In summary, our manuscript describes the following contributions:

- STARC-9 dataset with 630,000 high-quality tiles across nine tissue types for model training
- Stanford (independent from STARC-9) and TCGA-CRC tile-level validation datasets
- Domain-specific feature extractor based on a custom-trained autoencoder
- Code repository for DeepCluster++ for generating datasets from any WSI
- All models trained on the STARC-9 dataset

2 Related Works

Publicly available tile-level CRC Datasets: Despite the growing interest in computational pathology, there are relatively few publicly available H&E-stained tile-level CRC datasets for multi-class tissue classification. A significant contribution was made in [16] with a dataset containing 5,000 non-overlapping 150×150 pixel image tiles across eight tissue categories. This was later expanded to 100,000 tiles in the NCT-CRC-HE-100K dataset [15], with 224×224 pixel patches covering nine tissue types from 86 WSI and an additional 7,180 images from 50 WSI for the validation set. These datasets have been widely adopted for various downstream tasks such as tissue classification [33], segmentation [4], and MSI prediction [43]. Recently, the HMU-GC-HE-30K dataset [24] was released, containing 30,000 224×224 pixel patches of gastric cancer with detailed tumor microenvironment (TME) annotations. While a few additional CRC datasets exist online [36], many of them either lack direct public access or do not provide comprehensive annotations or tissue-level labels. In contrast, datasets like TCGA-COAD/READ [40] provide only unannotated WSI, requiring manual tile extraction for machine learning applications.

Existing methodologies for building histopathologic image datasets: Manual annotation (e.g., using QuPath [2]) of regions of interest (ROI) is slow and subjective and tends to favor easy regions, making it difficult to capture rare morphologies and maintain class balance as WSI size and complexity increase [31], [19]. Random sampling is susceptible to sampling error, wherein rare but clinically important morphologies are frequently missed, yielding imbalanced representations of tissue heterogeneity across WSI. Similarly, deep clustering (e.g., k-means on transfer-learned features [5], [9]) automates cluster formation, but sampling near cluster centroids biases toward common morphologies and under-represents intra-class variability required for robust supervised learning. Furthermore,

active learning [20] improves diversity by targeting model-uncertain samples, but requires iterative labeling and a seed of pre-labeled data.

Research gaps identified: Multi-class tissue classification for histopathology requires balanced, morphologically diverse datasets free of non-representative and low-quality tiles. However, existing datasets like NCT-CRC-HE-100K suffer from JPEG compression artifacts [11], and HMU-GC-HE-30K includes non-representative (e.g., incorrectly classified) tiles, leading models to learn spurious features. Similarly, TCGA-derived datasets exhibit sampling disparities and staining batch effects that affect model accuracy [18]. While techniques like cross-entropy uncertainty and probabilistic local outlier detection [39] can improve label quality, no cohesive pipeline exists for large-scale, balanced dataset curation. This limitation impacts downstream task performance, as observed in our initial experiments where models trained on HMU-GC-HE-30K and NCT-CRC-HE-100K achieved less than 90% accuracy, reducing classification effectiveness on the independent Stanford dataset. In addition, existing dataset construction methods are slow, biased, and fail to capture rare morphologies effectively.

3 DeepCluster++ for STARC-9 Construction

To address the limitations of existing CRC datasets, we developed a semi-automated framework, DeepCluster++ (Figure 2), to construct the STARC-9 dataset with 630,000 tiles across nine tissue types (ADI, LYM, MUS, FCT, MUC, NCS, BLD, TUM, NOR) shown in Figure 1, from over 200 WSI (patient demographic details are provided in Technical Appendices Section A) representing a diverse morphologic spectrum of CRC surgical resection specimens. This approach integrates unsupervised feature extraction, clustering to group similar tiles, equal-frequency binning for tissue diversity, and an expert verification phase, resulting in the creation of a high-quality dataset for downstream tasks such as classification, tumor segmentation, and prognostication.

3.1 Phase 1: Autoencoder Training

The first phase of our framework involves training an autoencoder (AE_CRC) to learn domain-specific feature representations from histopathologic tiles. Autoencoders are unsupervised learning models that encode input images into low-dimensional latent vectors through a convolutional encoder, then reconstruct them via a symmetrical decoder. This process forces the encoder to capture compact, informative features which preserve critical structural and visual details. While autoencoders have been used for small grayscale image collections [12], they have not been used to maximize diversity during tissue sample selection. For training AE_CRC, we sampled 100,000 tiles of size 256×256 pixels from 10 representative WSI (5 tumor and 5 normal) independent of the STARC-9 training and validation sets, covering all nine histologic tissue types. Tile extraction was performed using histogram-based thresholding at a 32 down-sample factor with a 25% tissue threshold to retain sparse tissues like ADI and MUC. Tile preprocessing included artifact removal and blank tile exclusion to create a high-quality ground truth set. Data augmentation techniques such as random rotations, flips, affine transformations, color jittering, and Gaussian blur were used to increase morphologic variability. The encoder consists of six convolutional layers with batch normalization and Leaky ReLU activations, producing a 32,768-dimensional latent vector. The decoder mirrors this architecture, using deconvolutional layers with a sigmoid activation in the final layer to reconstruct the input image. The AE_CRC was trained using a structural similarity index (SSIM) loss function (see Technical Appendices Section C for details), which captures structural features crucial for histopathology. The reconstruction quality of AE_CRC (as shown in Figure 2(a) for NCS, NOR, LYM) was checked to ensure that the autoencoder learned a representation of diverse histologic patterns. We chose a custom autoencoder because its domain-specific, reconstruction-driven features produce finer morphology-sensitive embeddings with lower compute requirements than broad foundation and pretrained models, yielding more coherent clusters and better prototypical and edge-case coverage (see Technical Appendices Section B for details).

3.2 Phase 2: Clustering and Sampling Tiles

We used only the frozen encoder of AE_CRC to generate unsupervised embeddings for clustering candidate tiles in each WSI. These embeddings served solely to guide morphology-aware tile sampling; they were not fed into any downstream classification or segmentation applications. Let the

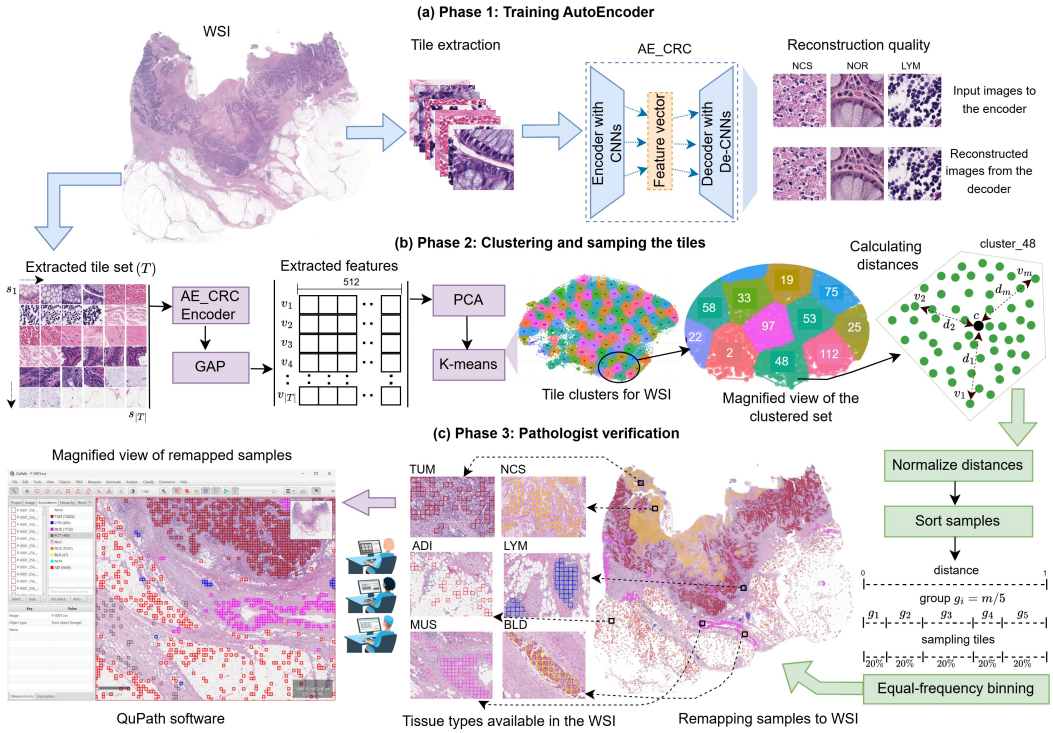


Figure 2: DeepCluster++ framework (Phases 1 and 2) followed by pathologist verification (Phase 3).

extracted tile set for a WSI be $T = \{s_1, s_2, \ldots, s_{|T|}\}$, in which each tile s_i was preprocessed as in Phase 1 and passed through the encoder to generate a 32,768-dimensional latent vector v_i. To reduce computational complexity and improve clustering performance, global average pooling (GAP) was applied to compress the feature vector to 512 dimensions, as shown in Figure 2(b). We subsequently applied principal component analysis (PCA) to further reduce the latent dimensionality to 256, thereby decreasing computational complexity and improving efficiency. These feature vectors were then clustered using the K-means algorithm [5], as in [32], to group tiles with similar morphology. We set the number of samples per cluster (m) to 400 to balance tissue diversity and representation quality, based on our empirical evaluation (see Technical Appendices Section G for details), as higher values (e.g., 800) risk including mixed tissue types, while lower values (e.g., 100) may reduce morphologic variation. Additionally, K-means was preferred over methods like DBSCAN, which lacks consistent cluster sizing. This approach also preserved local morphologic coherence, as adjacent clusters often contained similar tissue types, facilitating efficient sampling of diverse tissue patterns. The next step involved sampling tiles from each cluster to preserve morphologic diversity. For each cluster (e.g., cluster_48), we first computed the cluster centroid c and calculated the Euclidean distance from centroid for each tile as $d_i = \|v_i - c\|$, as shown in Figure 2(c). These distances were normalized to the range $[0, 1]$ to ensure consistency across clusters of varying densities. Tiles were then sorted by distance, and equal-frequency binning was applied to divide the samples into five distance-based groups ($g = 5$). This approach ensures that each group contains an equal number of tiles, preventing over-representation of dense regions near the cluster center and capturing a broad range of tissue patterns. Unlike equal-width binning, which often leads to imbalanced groups, this method maintains a uniform distribution of samples from near-centroid (homogeneous) to edge-of-cluster (diverse) tiles. Increasing (e.g., to 10) and decreasing (e.g., to 2) the number of bins respectively enhances and reduces variation, offering flexibility based on downstream requirements.

We sampled 20% of the tiles from each bin to ensure a comprehensive representation. The sampled tiles were then stored in separate class folders based on tissue type. Because these clusters did not carry semantic labels (e.g., "TUM" or "LYM"), we manually reviewed each output cluster to identify which tissue type it best reflected. To associate clusters with particular tissue types, once a cluster was labeled (for example, cluster_48 in Figure 2(b) was confirmed as "TUM"), we used the embedding space proximity to find adjacent clusters, such as 2, 97, 53, and 112, sharing similar feature representations implying similar histologic patterns. In our experiments, these neighboring

clusters consistently contained the same tissue morphology, allowing us to extend the "TUM" label across additional clusters with minimal additional time and effort. This local continuity within the feature space enabled more efficient exploration and sampling of tissue diversity. By iterating through this process (labeling a seed cluster and propagating its label to nearby clusters), we efficiently mapped appropriate clusters to the nine target tissue classes with modest manual effort.

3.3 Phase 3: Pathologist Verification and Final Dataset Assembly

The samples collected for all tissue types were mapped back to their original WSI location using QuPath [2] (Figure 2(c)) for pathologist verification of tissue-type classification. To support robust multi-tissue type classification, we fixed the number of tiles per tissue type at 70,000, resulting in a final STARC-9 dataset of 630,000 high-quality samples. While this decision might be debated, it was necessary for the consistent evaluation of 21,000 internal WSI. However, for other real-world applications, the number of tiles per class can be adjusted based on available data and downstream task requirements. All images were reviewed for classification accuracy by board-certified pathologists with subspecialty expertise in gastrointestinal (GI) pathology. In total, three board-certified GI pathologists participated in the review process: two pathologists with 13 and 41 years of experience, respectively, evaluated subsets of the dataset, while a third pathologist with 15 years of experience conducted a comprehensive final review of the entire dataset comprising 630,000 images.

4 Experiments and Results

Dataset description: STARC-9 is a comprehensive multi-tissue classification dataset consisting of 630,000 high-resolution non-overlapping 256×256 pixel tiles extracted from 40x magnification (0.25 micrometers/pixel) WSI. It includes nine clinically relevant tissue types: ADI, LYM, MUS, FCT, MUC, NCS, BLD, TUM, and NOR, capturing diverse, fine-grained tissue morphologies. To facilitate rigorous validation of models trained on STARC-9, two independent validation sets were prepared. (i) STANFORD-CRC-HE-VAL-SMALL contains 18,000 tiles (2,000 per tissue type) obtained from 20 WSI separate from the cases used in STARC-9 and was used for preliminary model testing, yielding 79.59% and 85.9% accuracy for models trained on NCT-CRC-HE-100K (NCT) [15] and HMU-GC-HE-30K (HMU) [24], respectively. This highlighted the need for larger, more diverse training sets, as these models struggled with mixed tissue-type tiles, achieving only 60% overall per-WSI tissue mapping accuracy. (ii) The primary validation set, STANFORD-CRC-HE-VAL-LARGE, includes 54,000 tiles (6,000 per class) from 50 WSI independent from STARC-9 and STANFORD-CRC-HE-VAL-SMALL for performance evaluation. Training (STARC-9) and validation (STANFORD-CRC-HE-VAL-SMALL and STANFORD-CRC-HE-VAL-LARGE) datasets were drawn from different patients, with no overlap between any of the datasets. Additionally, an external CURATED-TCGA-CRC-HE-VAL-20K dataset was prepared with 20,000 tiles extracted from 30 TCGA-CRC WSI to assess the robustness and generalizability of models trained on STARC-9. During benchmarking, models trained on STARC-9 (630,000 tiles, 9 classes), NCT (100,000 tiles, 9 classes), and HMU (30,000 tiles, 8 classes) were validated on seven overlapping tissue types (ADI, LYM, MUS, MUC, NCS, TUM, and NOR), ensuring fairness in performance evaluation.

Model description: To evaluate the utility of STARC-9, we conducted a series of benchmarking experiments using a diverse set of deep learning models, including baseline CNNs, SOTA transformer models, and pathology-specific foundation models. The objective was to assess classification performance, generalizability, and practical utility compared to models trained (fine-tuned) on publicly available datasets like NCT and HMU. Baseline models included ResNet-50, EfficientNet-B7, KimiaNet, and ViT-base, while SOTA models included DeiT-B, Swin Transformer-Base, and ConvNeXT-Base. Pathology-specific foundation models such as CTransPath, HiPT, Prov-Gigapath, Path-DINO, CONCH, UNI, Virchow, and VIM4Path were also tested to assess their generalizability on diverse tissue morphologies. Each model was trained on the STARC-9, NCT, and HMU datasets with Macenko normalization [26]. All models were fine-tuned with a batch size of 32, learning rate of 0.0001, weight decay of 1e-5, Adam optimization, and data augmentation (horizontal/vertical flips, random rotation, and color jittering) for 10 epochs. Maintaining identical training configurations, including batch size and optimizer settings, across datasets was important for a fair and unbiased model comparison. As our primary objective was to isolate the impact of the dataset on model performance, we kept the training configurations consistent to avoid possible confounding introduced by different hyperparameters being applied to the datasets. Model performance was evaluated using

precision, recall, macro F1 score, accuracy, and the number of trainable parameters. STARC-9-trained models consistently outperformed models trained on other datasets, exhibiting better generalizability.

Resource description: All experiments were conducted on the following platforms: (i) a local server with 8x NVIDIA L40S 48GB GPUs, and the Stanford (ii) Carina [35] and (iii) Marlowe high-performance computing platforms [14].

4.1 Multi-Class Tissue Classification

All models trained on STARC-9 demonstrated exceptional performance on the STANFORD-CRC-HE-VAL-LARGE validation dataset (Table 1). Among the baseline models, EfficientNet-B7 trained on STARC-9 achieved the highest overall accuracy (98.80%), with a 14.7% improvement over the best model trained on NCT (84.25%) and an 8.6% improvement over the best model trained on HMU (90.29%). In the SOTA category, Swin Transformer (Swin Trans-base) trained on STARC-9 achieved 98.79% accuracy, a 16.1% improvement over ConvNeXT-base trained on NCT (82.82%) and a 6.9% improvement over Swin Trans-base trained on HMU (91.88%). Among the pathology-specific foundation models, CTransPath trained on STARC-9 with 87M parameters achieved 99% accuracy, significantly outperforming UNI trained on NCT (80.43%) and HiPT trained on HMU (91.99%), emphasizing the importance of domain-specific pretraining. Custom models trained from scratch, such as a CNN and Histo-ViT trained on STARC-9, achieved accuracies of 97.81% and 96.32%, respectively, highlighting the ability of high-quality, domain-specific training data to enable effective representation learning without the overhead of pre-training and risk of overfitting. Overall, these results emphasize the importance of diverse, high-quality training samples for developing robust tissue classification models. The consistent improvements in precision, recall, and F1-macro scores across all tissue types highlight the advantage of STARC-9's data diversity, which contributed to the over 97% accuracy, even for models without extensive pretraining. CTransPath trained on STARC-9 consistently outperformed ViT-base (trained on NCT) and HiPT (trained on HMU) across all evaluation metrics on external validation sets.

Table 2 reports the precision, recall, F1-macro, and accuracy metrics for the top-performing models (with respect to accuracy in Table 1), when validated on STANFORD-CRC-HE-VAL-SMALL, STANFORD-CRC-HE-VAL-LARGE, and CURATED-TCGA-CRC-HE-VAL-20K. These top-performing models were: ViT-Base trained on NCT, HiPT trained on HMU, and CTransPath trained on STARC-9. For STANFORD-CRC-HE-VAL-SMALL, the STARC-9-trained CTransPath model achieved 99.75% precision, 99.73% recall, 99.74% F1- macro, and 99.73% accuracy - significantly higher than the other models, which showed lower recall and F1-macro scores. Similarly, on STARC-9-HE-VAL-LARGE, CTransPath maintained its lead, with 99.34% precision, 99.00% recall, 99.16% F1-macro, and 99.00% accuracy. Even on the more challenging STANFORD-TCGA-CRC-

Table 1: Multi-class tissue classification performance of baseline, SOTA, pathology foundation, and custom models trained on HMU, NCT, and STARC-9 for seven common tissue types (ADI, LYM, MUS, MUC, NCS, TUM, NOR) evaluated on the STANFORD-CRC-HE-VAL-LARGE dataset. The highest accuracy models for each dataset are highlighted in bold.

Model	Precision			Recall			F1-macro			Accuracy			No. of
	NCT	HMU	STARC-9	NCT	HMU	STARC-9	NCT	HMU	STARC-9	NCT	HMU	STARC-9	params.
Baseline models													
ResNet50 [10]	84.08	87.81	98.92	62.59	85.71	98.64	63.17	86.00	98.78	62.59	85.71	98.64	24 M
EfficientNet-B7 [38]	89.99	88.65	99.11	82.47	87.45	98.80	84.55	87.87	98.95	82.47	84.45	98.80	64 M
ViT-base [8]	92.71	91.57	98.49	84.25	90.29	98.09	87.30	90.87	98.28	**84.25**	90.29	98.09	86 M
SOTA models													
DeiT-B [41]	94.28	90.97	98.99	81.63	90.05	98.65	85.35	90.40	98.81	81.63	90.05	98.65	86 M
Swin Trans-base [21]	90.11	93.17	99.09	79.05	91.88	98.80	82.52	92.46	98.94	79.05	91.88	98.79	87 M
KimiaNet [32]	87.25	88.60	99.03	71.53	86.67	98.72	71.53	87.04	98.87	68.69	86.67	98.72	7M
ConvNeXT-base [22]	91.95	92.09	99.01	82.82	91.07	98.36	85.56	91.50	98.68	82.82	91.07	98.36	88 M
Pathology foundation models													
CTransPath [44]	90.11	93.17	99.34	79.05	91.88	99.00	82.52	92.46	99.16	79.05	91.88	**99.00**	87 M
HiPT [6]	90.92	93.21	98.64	74.51	91.99	98.32	77.41	92.54	98.47	74.51	**91.99**	98.32	86 M
ProvGigPath [45]	89.43	91.47	98.74	74.18	90.60	98.37	78.40	90.92	98.55	74.18	90.60	98.37	305 M
PathDino [1]	92.93	91.19	98.67	77.35	89.64	98.37	81.71	90.22	98.51	77.35	89.64	98.37	22 M
CONCH [25]	91.53	91.41	98.56	75.69	90.02	98.19	78.08	90.52	98.37	75.69	90.02	98.19	87 M
UNI [7]	94.55	93.03	98.67	80.43	91.80	98.25	84.42	92.36	98.45	80.43	91.80	98.26	88 M
Virchow [42]	92.51	92.35	98.63	79.02	91.23	98.28	82.05	91.69	98.45	79.02	91.23	98.28	305 M
VIM4PATH [28]	91.51	92.66	98.53	75.41	91.50	98.27	79.10	92.01	98.40	75.41	91.50	98.29	86 M
Customized models (trained from scratch)													
CNN	83.97	78.45	98.10	64.21	68.10	97.81	68.12	66.39	97.93	64.21	68.10	97.81	3.9 M
Histo-ViT	86.17	76.45	96.88	69.48	67.16	96.32	72.01	67.77	96.52	69.48	67.16	96.32	86 M

Table 2: Multi-class tissue classification performance of the best-performing models trained on HMU, NCT, and STARC-9 for seven common tissue types on the validation sets.

Validation dataset	Precision			Recall			F1-macro			Accuracy		
	NCT	HMU	STARC-9	NCT	HMU	STARC-9	NCT	HMU	STARC-9	NCT	HMU	STARC-9
STANFORD-CRC-HE-VAL-SMALL	88.52	90.22	99.75	76.19	88.34	99.73	79.34	89.16	99.74	76.19	88.34	**99.73**
STANFORD-CRC-HE-VAL-LARGE	92.71	93.21	99.34	84.25	91.99	99.00	87.30	92.54	99.16	84.25	91.99	**99.00**
CURATED-TCGA-CRC-HE-VAL-20K	89.69	92.21	99.03	72.42	90.9	98.85	76.74	91.45	98.94	72.42	90.9	**98.85**
IMP-CRS10K	63.29	65.06	96.70	42.77	61.99	94.88	45.85	62.46	95.55	69.62	76.40	**96.61**

HE-20K set, CTransPath consistently achieved near-perfect precision (99.05%), recall (98.88%), and F1-macro (98.96%), demonstrating excellent generalization and robustness across diverse tissue types. To further evaluate the generalizability of the model trained on STARC-9, we curated a small test set of the seven common tissue classes taken from 10 WSI from the IMP-CRS10K biopsy/polypectomy dataset [29]. In total, 1,093 image tiles were annotated for model performance validation, in which the STARC-9-trained model achieved a 95.55% F1-macro and 96.61% accuracy, consistently outperforming the HMU and NCT–trained models. It would also be interesting to evaluate the performance of the models trained on STARC-9 on the NCT and HMU datasets. However, as noted in Section 2 (and in reference [11]), the publicly available validation sets from NCT and HMU contain a substantial fraction of artifact-laden or mislabeled tiles, as well as "ambiguous" tiles with more than one tissue type represented within the same tile, despite only a single tissue-type label being assigned to the tile. In order to utilize these two datasets as reliable validation datasets, pathologist re-verification and correction/refinement of the tile-level labels would be necessary, which is labor-intensive and infeasible, given that the original WSI used to generate these two datasets were not publicly available for verification of label accuracy.

Feature map visualization analysis: Figure 3 illustrates the significant impact of training data quality on model feature activations for multi-class tissue classification. The figure presents activation maps generated by models trained on HMU, NCT, and STARC-9 for three representative ground truth input tiles (NOR, TUM, and mixed TUM) in panels (a), (b), and (c), respectively. Models trained on STARC-9 consistently focused on diagnostically relevant histologic features, aligning closely with pathologist evaluation patterns, while those trained on NCT and HMU often activated less diagnostically relevant regions. In Figure 3(a), while all three models correctly predicted the normal (NOR) class, the model trained on HMU activated more dispersed, less relevant regions, reflecting its exposure to less-representative training data. The NCT-trained model captures some vague cellular architecture, but lacks comprehensive coverage of relevant structures. In contrast, the STARC-9 model accurately focuses on the regions critical for the diagnosis, demonstrating the impact of well-curated, diverse training samples on models' ability to capture subtle, diagnostically relevant histologic features. In Figure 3(b), for a tumor (TUM) tile, both HMU and NCT-trained models highlight broad, non-specific regions, missing critical cellular features necessary for precise tumor identification. However, the STARC-9-trained model effectively captures the full structural context of the tumor, aligning closely with the pathologist's focus on tightly packed, hypercellular regions typical of tumor tissue. In the challenging case of a mixed tissue-type tile containing both tumor (TUM) and necrosis (Figure 3(c)), the HMU-trained model incorrectly classifies the tile as containing necrosis (NCS) and the NCT-trained model correctly classifies it as tumor (TUM), but with poorly localized feature activations, indicating a less precise spatial understanding. In contrast, the model trained on STARC-9,

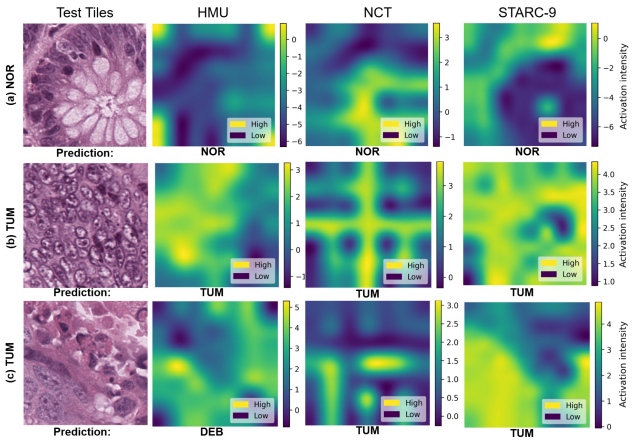


Figure 3: Feature map visualizations for the best models trained on HMU, NCT, and STARC-9.

which contains the complex, mixed tissue-type context often found in real-world WSI, accurately identifies the most clinically significant tumor regions. These feature map visualizations illustrate the high generalization capacity of models trained on STARC-9, further emphasizing the importance of diverse, high-quality training samples for robust, clinically relevant tissue classification.

Tissue map visualization: Figure 9 in Technical Appendices Section H shows tissue segmentation maps generated by remapping the tile-level classifications from models trained on STARC-9, NCT, and HMU back onto their respective WSI. This approach provides a quick, intuitive overview of WSI-level tissue composition for pathologist verification. For the sample regions highlighted for normal mucosa (NOR), necrosis (NCS), tumor (TUM), muscle (MUS), lymphoid tissue (LYM), mucin (MUC), and adipose tissue (ADI) in Figure 9(a), the model trained on STARC-9 consistently produced more accurate and contextually relevant predictions (Figure 9(d)), closely aligning with pathologist assessments. In contrast, models trained on NCT (Figure 9(c)) and HMU (Figure 9(b)) exhibited significant misclassification, particularly within challenging regions containing mixed tissue-type tiles. Notably, NCS classification was over 45% and 90% more accurate, when compared to the models trained on HMU and NCT, respectively. Additionally, blood-containing (BLD) regions, which were frequently misclassified as NCS by both the NCT and HMU-trained models, were correctly identified by the STARC-9-trained model (Figure 9(e)). Furthermore, the STARC-9-trained model demonstrated significantly lower confusion (over 80% error rate reduction) between the NOR, TUM, and MUC classes, compared to the models trained on HMU and NCT. Similarly, TUM regions, often misclassified as MUC by HMU and NCT (over a 30% error rate), were better delineated by the STARC-9-trained model. While all three models performed consistently across simple tissue types such as LYM and ADI, the STARC-9-trained model achieved over 85% accuracy on mixed tissue-type tiles, significantly outperforming the models trained on HMU (55%) and NCT (42%).

4.2 Tumor Tissue Segmentation

Among the most common downstream applications for multi-class tissue classification is tissue segmentation, especially tumor region segmentation, which allows for the automated identification and cropping of tumor-containing regions for subsequent annotation, ROI selection, diagnosis, and prognostication. This approach also facilitates downstream applications such as MSI [46] and other biomarker status prediction, and supports survival modeling [15] for risk stratification and personalized treatment planning. In this context, we conducted experiments to evaluate the effectiveness of models trained on HMU, NCT, and STARC-9 for tumor segmentation, focusing on their ability to accurately identify tumor regions that are important for clinical decision-making. As there were no publicly available CRC WSI repositories with readily usable, high-quality TUM masks for a larger scale experiment, and existing weakly-supervised tools did not provide the precision required for generating tissue segmentation masks, we prepared a test set by enlisting pathologists to manually annotate (ground truth) the TUM region in patches of size 2048×2048 pixels using QuPath [2]. We selected 45 patches (3 per slide) from 15 Stanford WSI and 50 patches (2 per slide) from 25 TCGA-CRC WSI, which were fully independent of our training and validation sets, for a more controlled evaluation of model performance [4]. Some patches contained mixed tissue types in order to evaluate the effectiveness of the trained models. For example, as shown in Figure 4, one region contained predominantly tumor, while the other included a mix of tumor and non-tumor tissue (NCS), providing more challenging segmentation.

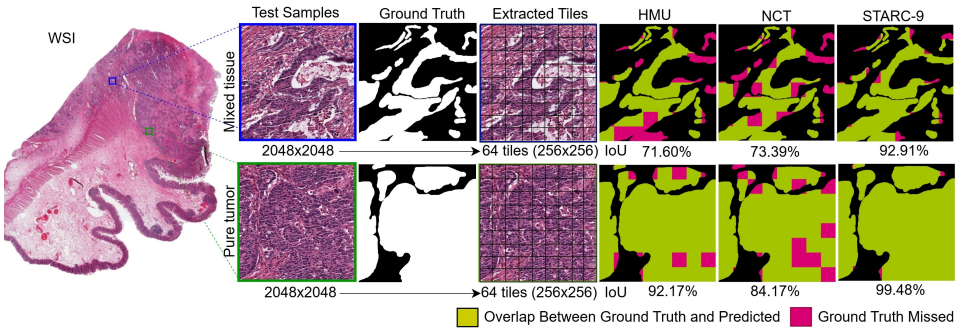


Figure 4: Tumor segmentation within 2048x2048 regions from a WSI from the CURATED-TCGA-CRC-HE-VAL dataset using tile-level classifiers trained on HMU, NCT, and STARC-9.

Table 3: Model evaluation for TUM segmentation on the Stanford and TCGA-CRC datasets.

Dataset	IoU (%)			Dice score (%)		
	NCT	HMU	STARC-9	NCT	HMU	STARC-9
Stanford	67.19 ± 21.53	64.68 ± 24.21	89.33 ± 8.76	78.20 ± 17.01	75.49 ± 21.01	90.47 ± 8.14
TCGA-CRC	51.94 ± 37.94	58.89 ± 29.42	88.81 ± 10.90	58.90 ± 31.38	68.85 ± 22.10	89.38 ± 9.14

For segmentation evaluation, each 2048x2048 pixel region was divided into 64 non-overlapping 256x256 pixel tiles and normalized to match the input requirements of the trained models. Tile-level classification was then performed using the best-performing model trained on each dataset. If a tile was correctly classified as TUM, its location within the ground truth segmentation mask was highlighted in green, while misclassified TUM tiles were marked in red, as shown in Figure 4. This approach allowed a direct visual comparison of each model's ability to accurately identify tumor regions. We observed that the model trained on STARC-9 significantly outperformed those trained on NCT and HMU, achieving an Intersection-over-Union (IoU) score of 92.91% for the mixed tissue-type sample, compared to 73.39% for NCT and 71.6% for HMU. This reflects the STARC-9-trained model's exceptional ability to capture fine-grained tissue features and effectively distinguish tumor regions, even within heterogeneous tissue contexts. For the pure tumor sample, the STARC-9-trained model also demonstrated higher performance, with a 99.48% IoU, significantly surpassing that of the models trained on NCT (84.17%) and HMU (92.17%). These results emphasize the critical role of diverse, high-quality training samples in developing robust, clinically relevant tissue classification models, particularly for challenging segmentation tasks.

Table 3 reports IoU and Dice scores for tumor segmentation on these held-out sets. Models trained on STARC-9 achieved mean Dice scores of 90.47±8.14% on the Stanford dataset and 89.38±9.14% on the TCGA-CRC dataset, approximately 14% and 17% higher than those trained on NCT and HMU when evaluated on the Stanford dataset, and 35% and 23% higher when evaluated on the TCGA-CRC dataset, respectively. Moreover, STARC-9-trained models exhibited substantially narrower standard deviations in both IoU and Dice scores, demonstrating more consistent and robust tumor delineation across diverse samples.

5 Conclusion

In this work, we introduce STARC-9, a large-scale, high-quality dataset for multi-class tissue classification for CRC histopathology. Comprising 630,000 non-overlapping high-resolution image tiles across nine clinically relevant tissue types, STARC-9 addresses critical limitations in existing datasets, including class imbalance, low tissue diversity, and low-quality tiles. We also present DeepCluster++, a flexible framework that combines unsupervised feature extraction, clustering, and equal-frequency binning to efficiently select diverse representative training examples from each WSI. Extensive benchmarking studies utilizing a wide range of deep learning models, including baseline CNNs, state-of-the-art transformers, pathology-specific foundation models, and custom deep learning models trained from scratch, demonstrate the superior classification performance of models trained on STARC-9 versus the publicly available NCT and HMU datasets, achieving over 98% accuracy on various independent validation datasets. The STARC-9-trained model also exhibited higher tumor segmentation accuracy, effectively capturing fine-grained tumor features critical for diagnosis and risk stratification, highlighting the importance of high-quality, diverse training data in model development.

Limitations and future scope: While STARC-9 contains extensive CRC tissue diversity across 9 tissue types, these may not exhaustively cover all potential tissue types found in CRC resections. Future work might focus on incorporating additional, more granular tissue classes, as well as expanding the dataset for multi-modal applications through the addition of large-scale image-caption pairs. Additionally, as STARC-9 is limited to CRC patients, its relevance for model validation for other cancer types not sharing similar tumor morphologies or background non-tumor tissue classes (for example, central nervous system tumors) remains to be explored. STARC-9 reflects local demographics, with limited Black and Native American representation. While race may not affect tissue morphology, broader inclusion is vital for fair, generalizable models. Lastly, we acknowledge that our dataset originates from a single institution and emphasize the need for future extensions incorporating multi-institutional data to enhance diversity and ensure fairness in downstream biomedical AI models.

Acknowledgments: Funding for this study was provided by the United States National Cancer Institute (NCI), National Institutes of Health (NIH) (R01 CA270437).

References

[1] Saghir Alfasly, Abubakr Shafique, Peyman Nejat, Jibran Khan, Areej Alsaafin, Ghazal Alabtah, and R. Tizhoosh, H. Rotation-agnostic image representation learning for digital pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[2] P. Bankhead, M. B. Loughrey, J. A. Fernández, and et al. Qupath: Open source software for digital pathology image analysis. *Scientific Reports*, 7:16878, 2017.

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[4] J. M. Bokhorst, I. D. Nagtegaal, F. Fraggetta, et al. Deep learning for multi-class semantic segmentation enables colorectal cancer detection and classification in digital pathology images. *Scientific Reports*, 13:8398, 2023.

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[6] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, June 2022.

[7] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] Zhibing Fu, Qingkui Chen, Mingming Wang, and Chen Huang. Whole slide images classification model based on self-learning sampling. *Biomedical Signal Processing and Control*, 90:105826, 2024.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Andrey Ignatov and Grigory Malivenko. Nct-crc-he: Not all histopathological datasets are equally useful. In *Computer Vision – ECCV 2024 Workshops*, pages 300–317. Springer Nature Switzerland, 2025.

[12] Mohamed Tarek Ismail, Hossam Sharara, Kareem Madkour, and Karim Seddik. Autoencoder-based data sampling for machine learning-based lithography hotspot detection. In *2022 ACM/IEEE 4th Workshop on Machine Learning for CAD (MLCAD)*, pages 91–96, 2022.

[13] Yiping Jiao, Junhong Li, Chenqi Qian, and Shumin Fei. Deep learning-based tumor microenvironment analysis in colon adenocarcinoma histopathological whole-slide images. *Computer Methods and Programs in Biomedicine*, 204:106047, 2021.

[14] Craig Kapfer, Kurt Stine, Balasubramanian Narasimhan, Christopher Mentzel, and Emmanuel Candes. Marlowe: Stanford's gpu-based computational instrument, January 2025.

[15] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal

cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):1–22, 01 2019.

[16] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M. Melchers, Lothar R. Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6:27988, 2016.

[17] Qi Ke, Wun-She Yap, Yee Kai Tee, Yan Chai Hum, Hua Zheng, and Yu-Jian Gan. Advanced deep learning for multi-class colorectal cancer histopathology: integrating transfer learning and ensemble methods. *Quantitative Imaging in Medicine and Surgery*, 15(3), 2025.

[18] F. Kheiri, S. Rahnamayan, M. Makrehchi, and Azam Asilian Bidgoli. Investigation on potential bias factors in histopathology datasets. *Scientific Reports*, 15:11349, 2025.

[19] Michał Koziarski, Bogusław Cyganek, Przemysław Niedziela, Bogusław Olborski, Zbigniew Antosz, Marcin Żydak, Bogdan Kwolek, Paweł Wąsowicz, Andrzej Bukała, Jakub Swadźba, and Piotr Sitkowski. Diagset: a dataset for prostate cancer histopathological image classification. *Scientific Reports*, 14:6780, 2024.

[20] Xiongquan Li, Xukang Wang, Xuhesheng Chen, Yao Lu, Hongpeng Fu, and Ying Cheng Wu. Unlabeled data selection for active learning in image classification. *Scientific Reports*, 14:424, 2024.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2022.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.

[23] Meng Lou and Yizhou Yu. Overlock: An overview-first-look-closely-next convnet with context-mixing dynamic kernels. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 128–138, 2025.

[24] S. Lou, J. Ji, H. Li, et al. A large histological images dataset of gastric cancer with tumour microenvironment annotation for ai. *Scientific Data*, 12:138, 2025.

[25] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[26] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, 2009.

[27] Eileen Morgan, Melina Arnold, A Gini, V Lorenzoni, C J Cabasag, Mathieu Laversanne, Jerome Vignat, Jacques Ferlay, Neil Murphy, and Freddie Bray. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from globocan. *Gut*, 72(2):338–344, 2023.

[28] Ali Nasiri-Sarvi, Vincent Quoc-Huy Trinh, Hassan Rivaz, and Mahdi S. Hosseini. Vim4path: Self-supervised vision mamba for histopathology images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6894–6903, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.

[29] P. C. Neto, D. Montezuma, S. P. Oliveira, D. Oliveira, J. Fraga, A. Monteiro, L. Ribeiro, S. Gonçalves, S. Reinhard, I. Zlobec, I. M. Pinto, and J. S. Cardoso. Imp whole-slide images of colorectal samples 2024, 2024. Data set.

[30] Xing-Liang Pan, Bo Hua, Ke Tong, Xia Li, Jin-Long Luo, Hua Yang, and Ju-Rong Ding. El-cnn: An enhanced lightweight classification method for colorectal cancer histopathological images. *Biomedical Signal Processing and Control*, 100:106933, 2025.

[31] Bálint Ármin Pataki, Alex Olar, Dezső Ribli, Adrián Pesti, Endre Kontsek, Benedek Gyöngyösi, Ágnes Bilecz, Tekla Kovács, Kristóf Attila Kovács, Zsófia Kramer, András Kiss, Miklós Szócska, Péter Pollner, and István Csabai. Huncrc: annotated pathological slides to enhance deep learning applications in colorectal cancer screening. *Scientific Data*, 9:370, 2022.

[32] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, Maral Rasoolijaberi, Milad Sikaroudi, Mohd Adnan, Sultaan Shah, Charles Choi, Savvas Damaskinos, Clinton JV Campbell, Phedias Diamandis, Liron Pantanowitz, Hany Kashani, Ali Ghodsi, and H.R. Tizhoosh. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70:102032, 2021.

[33] Ardhendu Sekhar, Ravi Gupta, and Amit Sethi. Few-shot histopathology image classification: Evaluating state-of-the-art methods and unveiling performance insights. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOIMAGING*, pages 244–253. INSTICC, SciTePress, 2024.

[34] Dai Shi. Transnext: Robust foveal visual perception for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17773–17783, June 2024.

[35] Stanford University. Carina computing platform. `https://carina.stanford.edu`, n.d. Accessed: 2025-10-15.

[36] Marie (Duc) Stettler. Histopathology datasets for machine learning. `https://github.com/marieduc/Histopathology-Datasets-for-Machine-Learning`, n.d. GitHub repository.

[37] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. Pathgen-1.6m: 1.6 million pathology image-text pairs generation through multi-agent collaboration, 2024.

[38] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML) / PMLR 97*, pages 6105–6114, 2019.

[39] Ishrat Zahan Tani, Kah Ong Michael Goh, Md Nazmul Islam, Md Tarek Aziz, S. M. Hasan Mahmud, and Dip Nandi. Addressing label noise in colorectal cancer classification using cross-entropy loss and plof methods with stacking-ensemble technique. *Applied Computational Intelligence and Soft Computing*, 2025(1):6552580, 2025.

[40] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487:330–337, 2012.

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

[42] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30:2924–2935, 2024.

[43] Wenyan Wang, Wei Shi, Chuanqi Nie, Weipeng Xing, Hailong Yang, Feng Li, Jinyang Liu, Geng Tian, Bing Wang, and Jialiang Yang. Prediction of colorectal cancer microsatellite instability and tumor mutational burden from histopathological images using multiple instance learning. *Biomedical Signal Processing and Control*, 104:107608, 2025.

[44] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

[45] H. Xu, N. Usuyama, J. Bagga, and et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024.

[46] Rikiya Yamashita, Jin Long, Teri Longacre, Lan Peng, Gerald Berry, Brock Martin, John Higgins, Daniel L Rubin, and Jeanne Shen. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *The Lancet Oncology*, 22(1):132–141, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 4 comprehensively covers the claims made in the abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It is included in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: It is included in the Supplementary Materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: It is included in the Supplementary Materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 includes all the necessary information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Since we fine-tune our models on the entire STARC-9 dataset rather than using repeated train/test splits, we do not perform cross-validation or multiple independent runs. Consequently, there are no mean-and-standard-deviation estimates to report, and error bars are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: It is discussed in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our STARC-9 dataset aims to improve colorectal cancer diagnostics, potentially enhancing patient outcomes, but may also pose risks if used in biased or non-clinical contexts without proper oversight.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The STARC-9 dataset is curated from clinical histopathology images, which have low risk for misuse as they are fully de-identified and intended solely for medical research and computational pathology applications. The dataset contains no identifiable patient information.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The STARC-9 dataset is built on original WSI collected under institutional review board (IRB) approval from Stanford University School of Medicine. All external datasets used for model benchmarking, including NCT-CRC-HE-100K (Kather et al., 2019) and HMU-GC-HE-30K (Lou et al., 2025), are appropriately cited clearly mentioned in the manuscript.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The STARC-9 dataset, introduced in this paper, is thoroughly documented with detailed metadata, including tissue type labels, tile extraction methods, preprocessing steps, and license information. The dataset is released under the CC-BY 4.0 license, with comprehensive documentation provided to ensure transparency, reproducibility, and proper usage in downstream computational pathology applications.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The STARC-9 dataset was curated from clinical histopathology images, which does not involve crowdsourcing or direct interaction with human subjects. All data used were fully de-identified and collected under appropriate institutional review board (IRB) approval, with waived informed consent.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The STARC-9 dataset was curated under IRB approval with waived informed consent, ensuring ethical compliance.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core methods in this research do not involve LLMs as an important or original component.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Technical Appendices

A Patient demographic details

For the STARC-9 dataset, as shown in Figure 5, 53% of patients were male and 47% female, with racial/ethnic distribution 64.5% White, 14% Hispanic, 12.5% Asian/Pacific Islander, 3.5% Black, and 5.5% Other/Unknown. Age range: 23-97 yrs (mean 62.9 yrs, standard deviation 16 yrs). Tumor grade distribution: 13% Grade 1 (n=26), 65% Grade 2 (n=130), 20% Grade 3 (n=4), and 2% Grade Not Applicable (n=4, all medullary carcinomas). Histologic subtypes: 88.5% (n=177) Adenocarcinoma, 6.5% (n=13) Mucinous Adenocarcinoma, 2% (n=4) Medullary Carcinoma, 1.5% (n=3) Signet-ring Cell Carcinoma, and 1.5% (n=3) Carcinoma, Type Undetermined. Regarding microsatellite instability status: 56% (n=112) microsatellite stable (MSS), 9% (n=18) microsatellite unstable (MSI), 35% MSI status unknown. Tumor location: 43% (n=86) Right/ transverse colon, 28% (n=56) Left colon/splenic flexure/rectosigmoid, and 29% (n=58) Rectum.

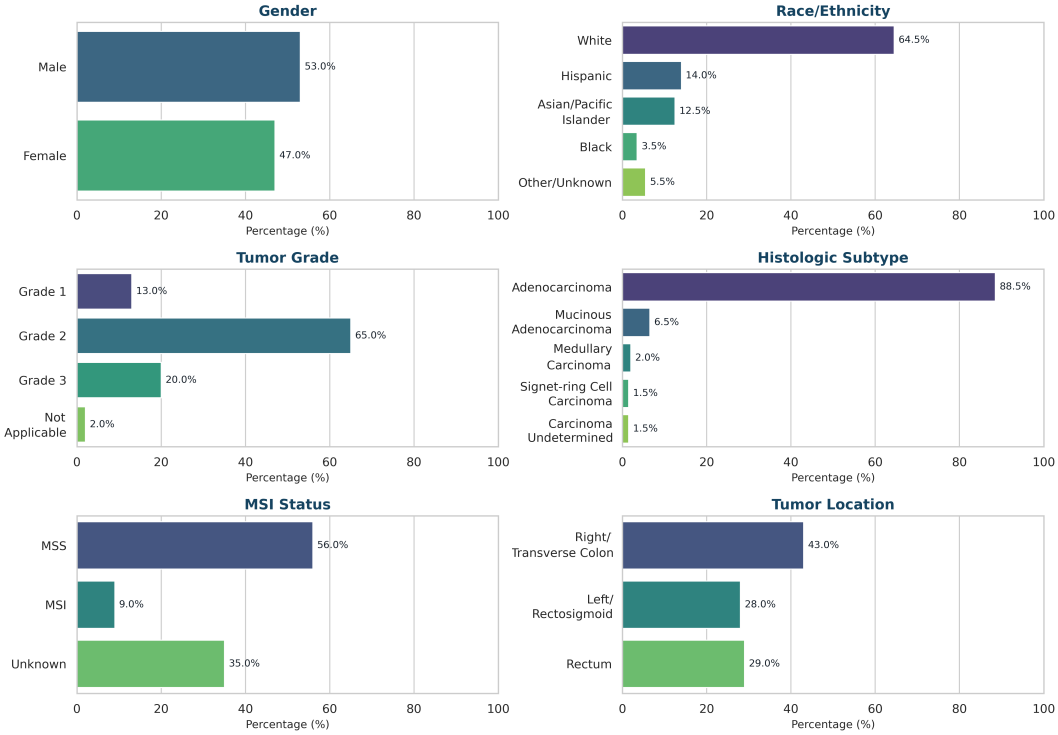


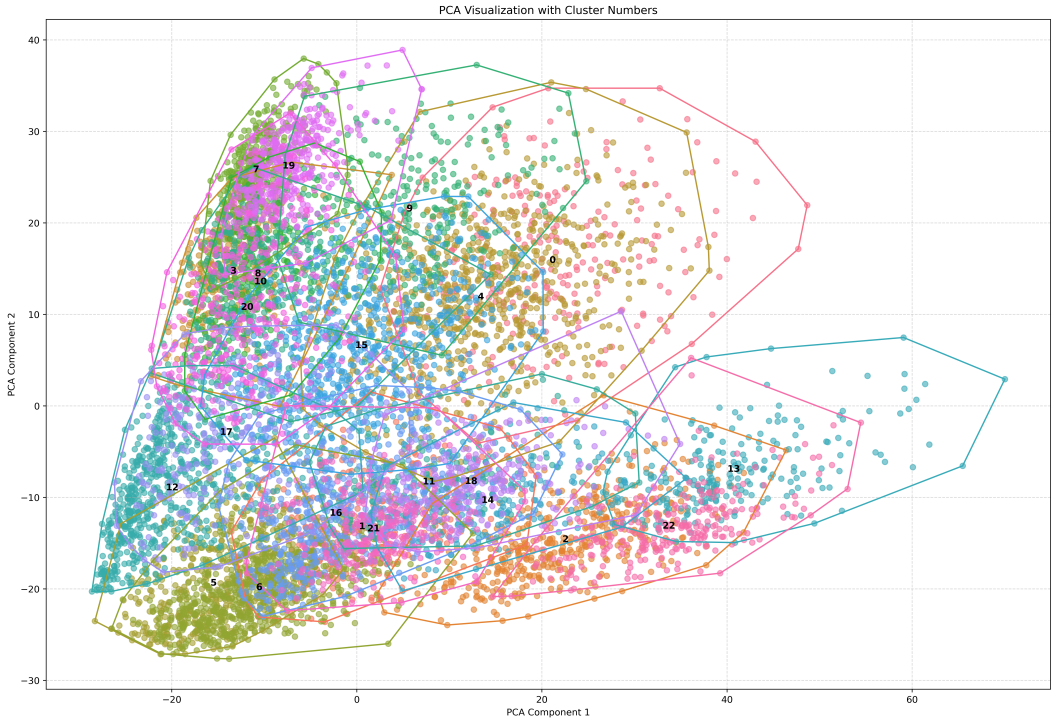
Figure 5: STARC-9 patient demographic details.

As WSI naturally vary in the amount and composition of tissue present on a slide, the per-patient and even per-tissue class tile counts were inherently imbalanced (for example, WSI with a high tumor volume tended to contain more NCS and MUC tiles). As this imbalanced tile distribution reflects the naturally diverse/heterogeneous tissue-type class distribution, it would be infeasible to balance the number of tiles of each tissue class per patient without running into the issues of (1) having an insufficient number of tiles from some patients and (2) needing to discard valid/informative tiles from some patients. Therefore, we did not seek to balance the number of tiles per patient.

B Advantages of using an AutoEncoder for feature extraction

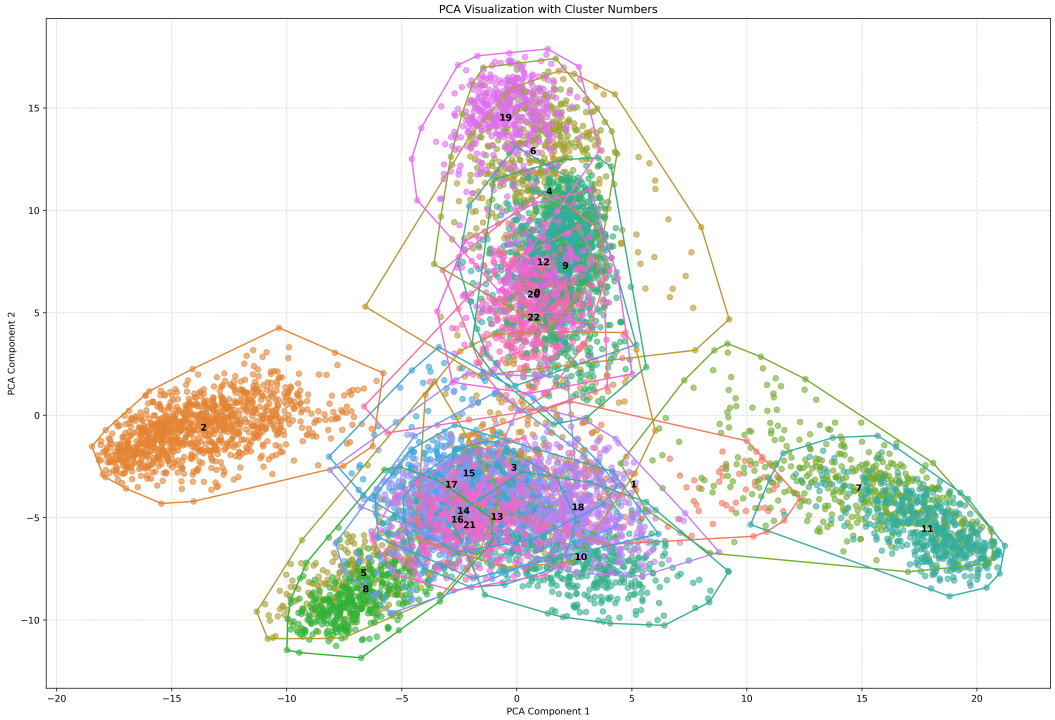
In constructing a high-quality, diverse, and representative histopathology dataset, the choice of a feature extractor is critical. We chose a custom-trained autoencoder (AE_CRC) over off-the-shelf pathology foundation models for three main reasons:

Task-specific features: To understand the role of encoders in sample selection, we analyzed 9,000 samples across nine tissue types from STARC-9 (1,000 per class). After feature extraction, embeddings were reduced to 256 dimensions and clustered using K-means (400 samples per cluster). As shown in Figure 6, supervised encoders such as, ResNet50 (trained on natural images) (Figure 6(a)) and KimiaNet (trained on images from pathology WSI) (Figure 6 (b)) exhibited scattered and

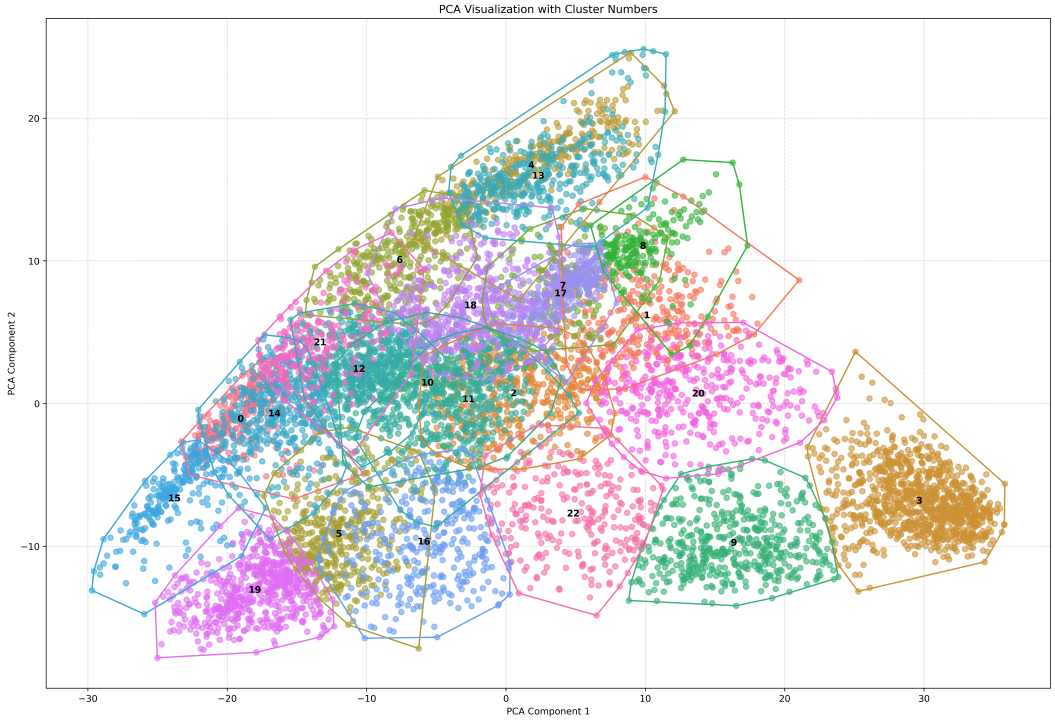


(a) ResNet50

(b) KimiaNet



(c) UNI



(d) AutoEncoder

Figure 6: Feature visualization of different encoders (a) ResNet50, (b) KimiaNet, (c) UNI, and (d) encoder from custom trained AutoEncoder, for 9000 samples of nine different tissue types.

overlapping representations among tissue types. Pathology foundation models are typically trained with classification or contrastive losses to distinguish major diagnostic categories. However, these discriminative objectives often fail to capture subtle intra-class variations essential for representing the full morphologic spectrum of tissue types. For example, despite UNI (Figure 6(c)) achieves strong self-supervised representations, they tend to over-separate biologically related tissues, reducing intra-class coherence.

These perform well on label-based separation but overlook fine-grained structural differences critical for interpretability and intra-class diversity. In contrast, our autoencoder-based encoder (Figure 6d(d)) learns structure-preserving, domain-specific representations by optimizing a reconstruction loss using the Structural Similarity Index (SSIM). This aligns clusters according to true morphologic proximity, preserving biologically meaningful relationships. This structure-preserving representation enables balanced sampling, helps exploring the surrounding clusters for similar tissue types, and supports the construction of morphologically diverse, clinically meaningful datasets such as STARC-9, where representational fidelity is critical for robust downstream model development.

Domain-specific sensitivity: By training exclusively on over 100,000 CRC tiles, AE_CRC becomes finely attuned to colorectal (and other tubular gastrointestinal tract) histopathology. It learns to distinguish the fine-grained features present within tissue types such as NOR, MUC, TUM, and NCS. This is in contrast to existing pathology foundation models, which are trained on a wide variety of organs and tasks, causing these models to overlook the specific fine-grained features necessary for accurate representation within the latent space.

Efficiency and scalability: Large foundation models (e.g., CTransPath, UNI, CONCH) require substantial GPU resources and slower embedding times. AE_CRC, in contrast, is lightweight and fast to implement on standard hardware, making it practical for clustering 630,000 tiles in DeepCluster++ without incurring prohibitive compute costs.

In our experiments, clustering with AE_CRC embeddings produced more coherent and morphology-driven groups compared to models trained on natural images (e.g., ImageNet) or contrastive learning-based pathology encoders. This allowed DeepCluster++ to effectively sample both prototypical and edge-case tiles, ensuring comprehensive coverage of histologic diversity. The reconstruction-based learning objective thus aligns well with the goal of building a large-scale, diverse histopathology benchmark dataset. However, when applying this framework to a different dataset, it may be necessary to retrain the AE_CRC model on the target data before integration into DeepCluster++.

C SSIM vs MSE loss functions in AutoEncoder

To ensure the AutoEncoder learned higher-level histologic and morphologic structures rather than low-level pixel statistics, we trained it using a structural similarity loss (SSIM). Compared to a model trained with mean squared error (MSE) loss, the SSIM-based AutoEncoder achieved significantly better reconstruction quality on the validation set STANFORD-CRC-HE-VAL-SMALL, showing lower pixel error (0.0012 vs 0.0015), higher SSIM (0.9262 vs 0.8863), and greater Peak Signal-to-Noise Ratio (PSNR) (32.48 dB vs 28.53 dB) on average. These metrics confirm high-fidelity reconstruction of complex tissue types such as necrosis (NCS) and tumor (TUM), as illustrated in Figure 2(a). By explicitly optimizing for texture, contrast, and spatial structure, SSIM encourages the encoder to capture perceptually meaningful features rather than minimizing pixel-level intensity differences. This results in a latent space rich in fine-grained morphologic representations, leading to distinct and coherent clusters in DeepCluster++. In contrast, MSE-based models often blur subtle variations and reduce cluster purity by focusing on pixel accuracy rather than structural integrity. The observed reconstruction quality demonstrates that the AutoEncoder-based feature extractor surpasses task-specific and contrastive encoders in capturing diverse and biologically relevant tissue morphologies, thereby offering a more reliable foundation for clustering and representation learning.

D Additional experiments with recent models

We trained and evaluated recent models such as TransNeXt [34], OverLoCK [23], and Beit-base [3] with masked image modeling on the NCT, HMU, and STARC-9 datasets. The comprehensive results are presented in Table 4. While these recent models demonstrate competitive performance, they did not surpass our best-performing combinations: CTransPath on STARC-9, HiPT on HMU,

Table 4: Evaluation of additional models on the validation datasets.

Model	F1-Macro (%)			Accuracy (%)			Model size
	NCT	HMU	STARC-9	NCT	HMU	STARC-9	
STANFORD-CRC-HE-VAL-SMALL							
TransNeXt [34]	59.68	56.57	98.60	73.89	64.22	98.61	110M
OverLoCK [23]	63.75	62.09	97.20	79.00	68.98	97.22	24.3M
Beit-base [3]	59.95	76.48	98.17	74.08	86.17	98.20	86.5M
STANFORD-CRC-HE-VAL-LARGE							
TransNeXt [34]	59.62	55.05	98.58	73.55	62.78	95.68	110M
OverLoCK [23]	63.69	55.56	98.85	79.29	64.09	98.87	24.3M
Beit-base [3]	62.01	78.11	98.61	77.03	88.40	98.68	86.5M
CURATED-TCGA-CRC-HE-VAL-20K							
TransNeXt [34]	58.95	51.92	95.57	72.65	61.70	96.07	110M
OverLoCK [23]	64.47	53.93	95.38	80.25	62.98	95.56	24.3M
Beit-base [3]	57.45	72.57	97.93	71.75	82.11	97.87	86.5M

and ViT-base on NCT (as reported in Table 1. Notably, models trained on STARC-9 consistently achieve superior performance across all architectures, validating our dataset's quality and diversity.

Regarding contrastive learning, we have extensively evaluated several state-of-the-art pathology-specific foundation models (CTransPath, UNI, CONCH, Virchow, etc.) that employ contrastive learning and are pre-trained on histopathologic images, as comprehensively discussed in Section 4.1. These experiments demonstrate that our STARC-9 dataset enables competitive performance even with the latest architectural advances.

E Relationship between model size and dataset-specific performance

Across the NCT, HMU, and STARC-9 datasets, the relationship between model size and classification accuracy did not follow a simple linear trend, as shown in Figure 7. Instead, model performance appeared to depend far more on architectural design and domain alignment than on the number of parameters. For instance, when trained on the STARC-9 dataset, CTransPath achieved the highest accuracy (99%), despite having only about 87 million parameters, which is considerably smaller than models such as ProvGigPath or Virchow (both 305 million parameters), as shown in Table 5. This result suggests that histopathology-specific pretraining and architectural efficiency enable CTransPath to capture subtle morphologic cues better than very large, general-purpose models that risk overfitting given the moderate dataset size of STARC-9.

When trained on the HMU dataset, HiPT outperformed all other architectures, with an accuracy of 91.99%. Like CTransPath, HiPT belongs to the class of medium-sized transformer models (around 86 million parameters). Its hierarchical patch-embedding design effectively integrates local and global contextual features, which seems particularly beneficial for tissue patterns requiring multiscale spatial reasoning. Larger models such as ProvGigPath and Virchow again offered no significant performance improvement, implying diminishing returns once models exceed roughly 100 million parameters.

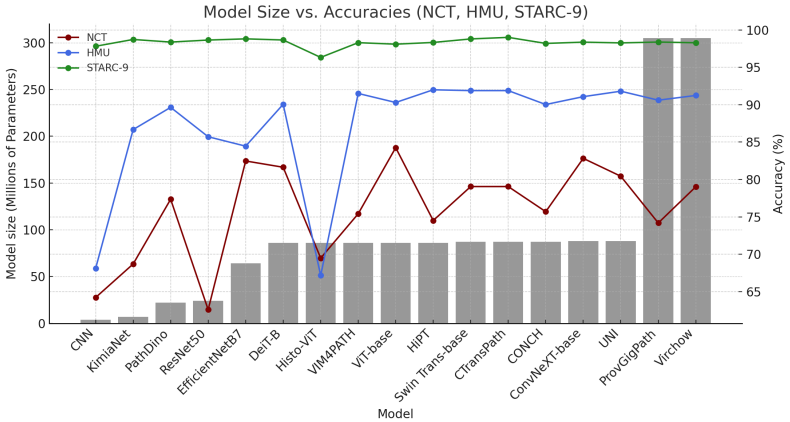


Figure 7: Relationship between model size and dataset-specific performance.

Table 5: Model size and performance (bold-face denotes highest performance on each dataset).

Model	No. of params.	NCT	HMU	STARC-9
CNN	3.9 M	64.21	68.1	97.81
KimiaNet	7M	68.69	86.67	98.72
PathDino	22 M	77.35	89.64	98.37
ResNet50	24 M	62.59	85.71	98.64
EfficientNet-B7	64 M	82.47	84.45	98.8
ViT-base	86 M	**84.25**	90.29	98.09
HiPT	86 M	74.51	**91.99**	98.32
Histo-ViT	86 M	69.48	67.16	96.32
VIM4PATH	86 M	75.41	91.5	98.29
DeiT-B	86 M	81.63	90.05	98.65
Swin Trans-base	87 M	79.05	91.88	98.79
CTransPath	87 M	79.05	91.88	**99**
CONCH	87 M	75.69	90.02	98.19
ConvNeXT-base	88 M	82.82	91.07	98.36
UNI	88 M	80.43	91.8	98.26
ProvGigPath	305 M	74.18	90.6	98.37
Virchow	305 M	79.02	91.23	98.28

The HMU dataset thus appears best served by architectures that balance representational power with generalization capacity rather than raw scale.

For the NCT dataset, ViT-base achieved the best performance (84.25%) among all evaluated models. Although it shares a similar parameter range with HiPT and CTransPath, ViT's pure attention mechanism captures patch-level variations and color normalization differences characteristic of the NCT slides. In contrast, smaller CNN-based models (e.g., a 3.9-million-parameter CNN) underperformed due to limited capacity for modeling long-range dependencies, while much larger networks did not yield further gains. This reinforces that the optimal model capacity for histopathology datasets often lies within a moderate range where sufficient complexity is achieved without overfitting risk.

When comparing the training results across all three datasets, a trend is consistent: medium-sized transformer architectures, typically between 80 and 90 million parameters, deliver the most reliable and generalizable performance. Larger models do not necessarily outperform smaller ones, as the marginal benefit of additional parameters diminishes once the representational capacity surpasses the diversity of the dataset. These findings emphasize that, for computational pathology, model design and domain pretraining (resulting in effective representation learning tailored to tissue morphology and staining variability) are far more important than model size.

F Advantages of DeepCluster++ for computational pathology

The proposed approach significantly reduces the manual burden of annotation and tile selection, compared to the conventional approach to constructing tissue-type classification datasets (which involves manual pathologist delineation of ROI within a WSI, followed by extraction of tiles from these ROIs). In contrast, with our automated DeepCluster++ framework, once the tiles have been collected for each tissue class within a WSI (which would normally require a significant amount of human time and effort using the conventional manual approach), a pathologist can simply use QuPath software to re-map the collected tiles back onto the original WSI from which they were taken, then confirm through a quick WSI-level visual inspection that the tiles for each tissue class were correctly assigned by DeepCluster++. This quality-control pass takes less than five minutes per slide, which is significantly less time than would be required to perform manual ROI annotation for each tissue class (following the conventional annotation and tile selection approach). By restricting the pathologist's workload to this final WSI-level verification step, our DeepCluster++ allows for the collection of tissue-type specific datasets with high-confidence labels and significantly reduced manual effort.

G Ablation study

For the DeepCluster++ framework, performing a comprehensive ablation study by varying multiple configuration parameters and extracting large-scale datasets from 200 WSI is time-intensive, as it

requires tile verification prior to downstream task evaluation. Therefore, we carried out a targeted ablation on 10 randomly selected WSI (independent of training/validation) for the TUM and NCS classes:

Number of samples per cluster *(m)*: Determining the number of clusters (K) and m is very challenging, and especially with histopathology-based tiles, it is difficult to set one value for the number of clusters and samples. We tested $m = 100$ to 800, finding that a small ($m \approx 100$) yielded many tiny, redundant clusters, while a large ($m > 800$) resulted in mixing of tissue types. We found that $m \approx 400$ balanced intra-cluster purity and inter-cluster diversity. Table 6 shows the trade-off between the size of the cluster and the purity (inverse of the Shannon entropy) based on the fixed number of significant variations in TUM and NCS tissue morphology. As m increases, purity decreases (entropy rises). We selected $m = 400$, as it yielded substantial morphologic variation while maintaining low tissue-type admixture, representing the best balance for our goal. If deciding m is complex for a dataset, it is recommended that the number of clusters and samples per cluster be set using $K = m = \sqrt{T}$ [37], where T is the number of tiles from a WSI.

Table 6: Cluster quality analysis.

m	**Average Entropy**	**Morphologic variation**
200	0.12	Low
400	0.41	High
600	1.73	High (but mixed tissue types)
800	2.17	High (but highly mixed tissue types)

Number of bins *(g)*: Fixing the number of tiles per cluster at $m = 400$ for a single tissue type, we performed an ablation on the number of equal-frequency distance bins, $g \in \{1, 2, ..., 10\}$. Each cluster contained approximately 4 to 6 distinguishable morphologic variants (e.g., structural subtypes within TUM). To assess within-bin homogeneity, we computed the average inverse normalized Shannon entropy, I(g), across bins, capturing the consistency of morphologic patterns within each bin. As illustrated in Figure 8, $I(1)$ was the highest because a single coarse bin merges all five morphologic variants, causing admixture of heterogeneous tiles and redundancy in sampled images. In contrast, $I(10)$ was the lowest, as the data became excessively fragmented, with similar patterns being split across different bins, with each bin containing few tiles with nearly identical appearances, reducing the overall morphologic diversity.

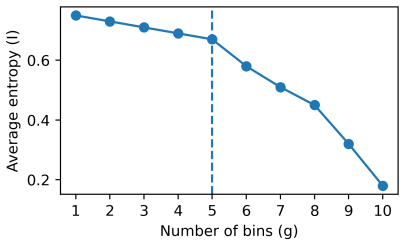


Figure 8: Ablation on the number of bins (g) using average inverse Shannon entropy.

Interestingly, the range $g = 4$ to 6 provided a balanced configuration: bins exhibited sufficient internal similarity while maintaining broad coverage across the morphologic continuum, from prototypical (near-centroid) to atypical (edge-of-cluster) tiles. In practice, the number of significant variants can differ across tissue types, making it impractical to fix g purely by empirical morphologic counts. Therefore, to mitigate excessive variability at smaller bin counts ($g = 4$) and prevent over-fragmentation at larger counts ($g > 5$), we identified $g = 5$ as the optimal trade-off between within-bin similarity and across-bin diversity. This configuration consistently preserved meaningful morphologic transitions while maintaining stable sampling performance across tissue types.

Sampling percentage (20%): Following the selection of g=5, we sampled an equal 20% of tiles from each bin to ensure a uniform and unbiased representation across the morphologic spectrum. This strategy guaranteed that all morphologic variants, from highly prototypical to rare atypical appearances, were proportionally included in the dataset. The 20% sampling rate offered a practical balance between computational efficiency and morphologic coverage. Depending on the specific requirements of downstream tasks (e.g., classification, segmentation, or survival modeling), the sampling ratio can be scaled up or down to expand or contract the dataset size while maintaining representational consistency.

H Tile-level prediction map overlaid on the WSI.

NCT: adipose (ADI), lymphocytes (LYM), smooth muscle (MUS), mucus (MUC), debris (DEB), colorectal adenocarcinoma epithelium (TUM), normal colon mucosa (NORM), cancer-associated stroma (STR), background (BACK)

HMU: adipose tissue (ADI), lymphocyte aggregates (LYM), muscle (MUS), mucus (MUC), debris (DEB), tumor epithelium (TUM), normal mucosa (NORM), stroma (STR)

STARC-9: adipose tissue (ADI), lymphoid tissue (LYM), muscle (MUS), fibroconnective tissue (FCT), mucin (MUC), necrosis (NCS), blood (BLD), tumor (TUM), and normal mucosa (NOR)

Names for the same tissue type in different datasets: DEB in NCT/HMU corresponds to NCS in STARC-9 and NORM in NCT/HMU corresponds to NOR in STARC-9.

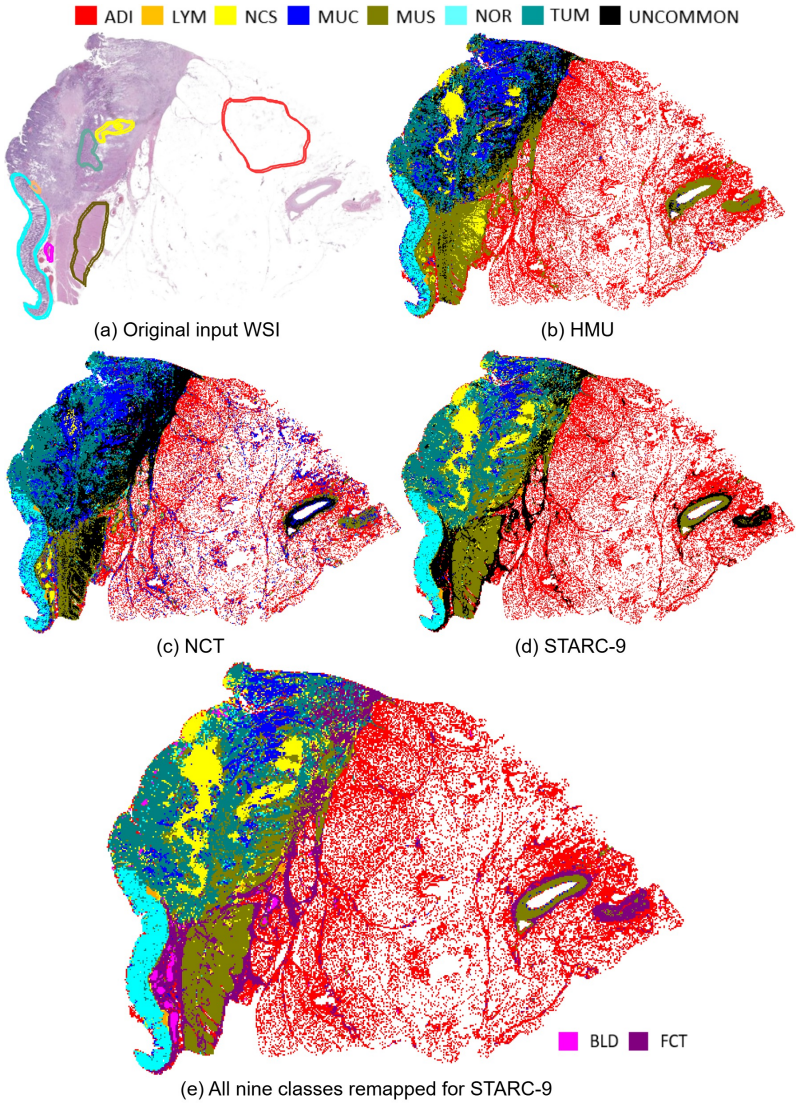
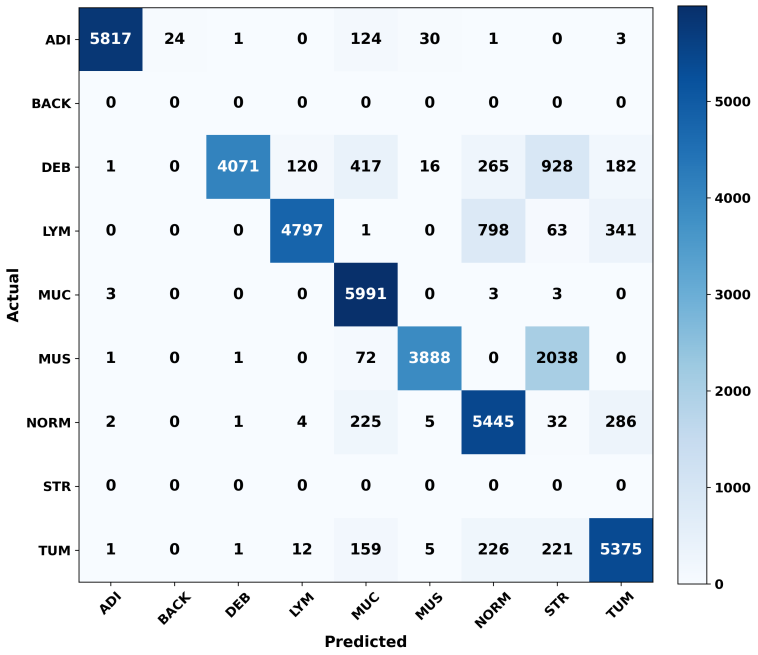
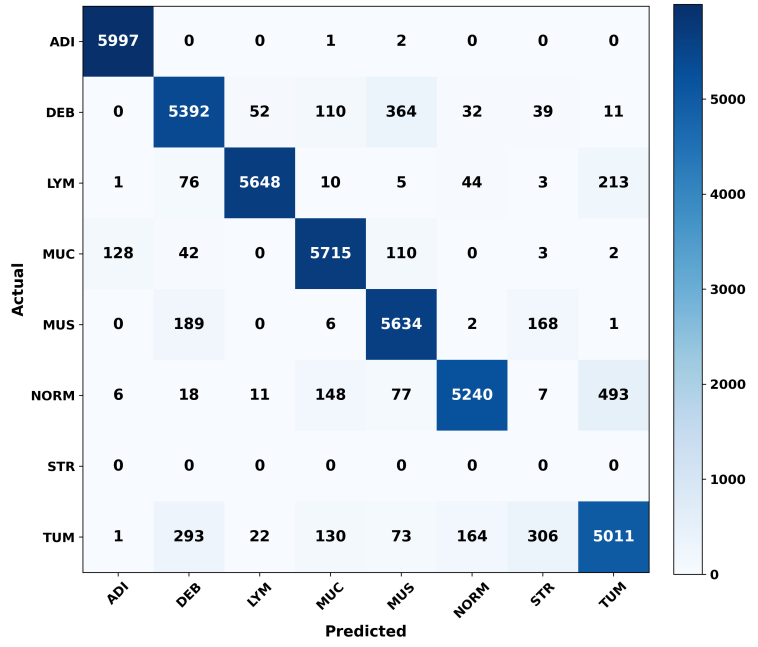


Figure 9: Tile-level prediction maps overlaid on a given input WSI (a) using the best-performing models trained on (b) HMU, (c) NCT, and (d) STARC-9 for the seven common tissue classes (ADI, LYM, MUS, MUC, NCS, TUM, NOR). Tiles assigned to classes outside these seven (e.g., stroma-STR in HMU/NCT and BLD and FCT in STARC-9) are shown in black (Uncommon tissue classes). Panel (e) shows all nine classes (included in STARC-9) mapped back to the input WSI.

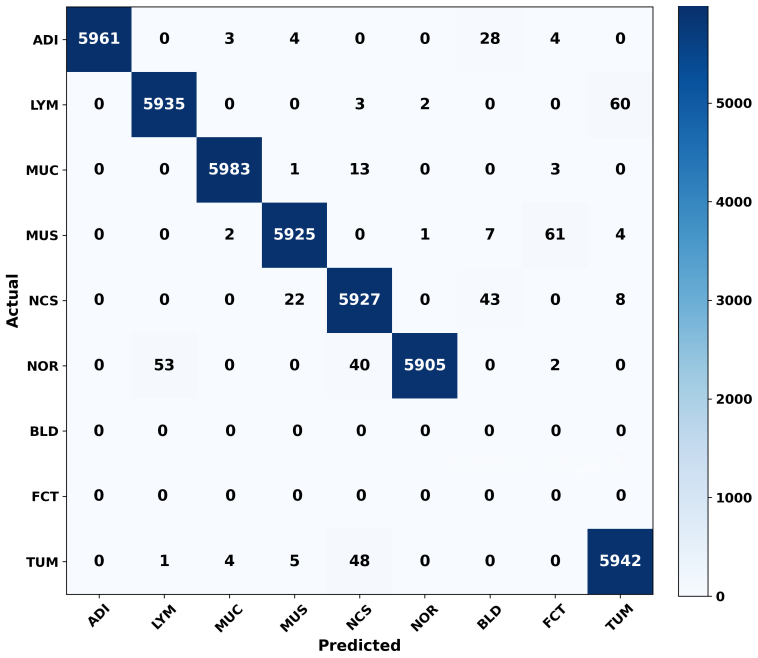
I Confusion matrices for the best-performing models (trained on NCT, HMU, and STARC-9) on STANFORD-CRC-HE-VAL-LARGE for seven common tissue types.



(a) NCT



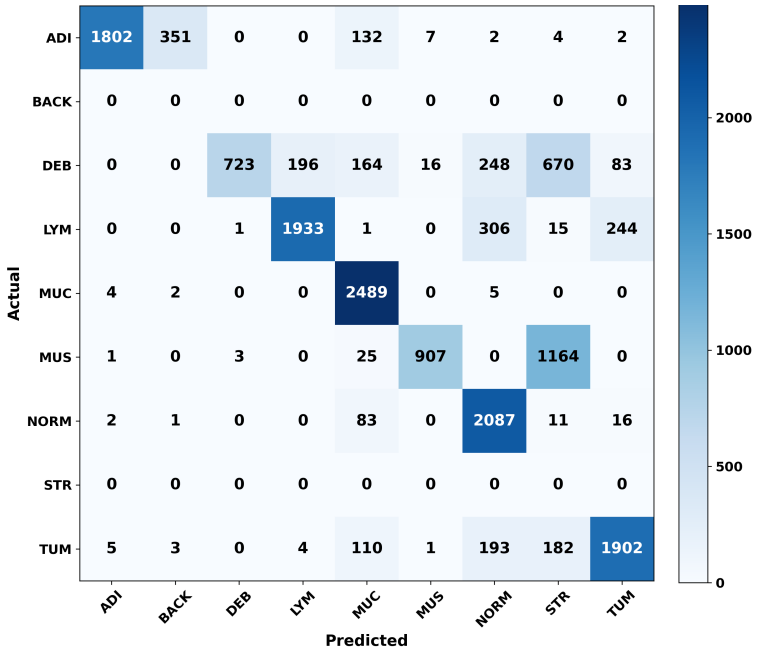
(b) HMU



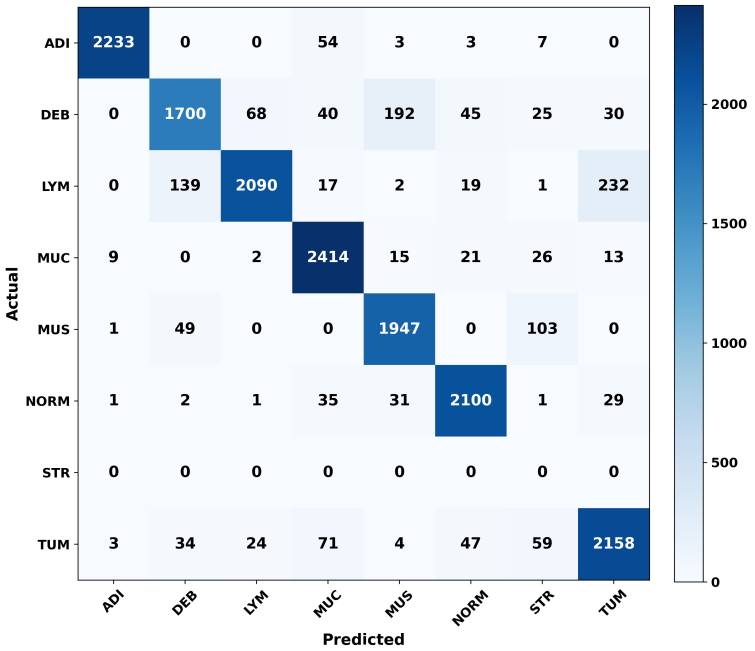
(c) STARC-9

Figure 10: Confusion matrices for the best-performing models on STANFORD-CRC-HE-VAL-LARGE for seven common tissue types.

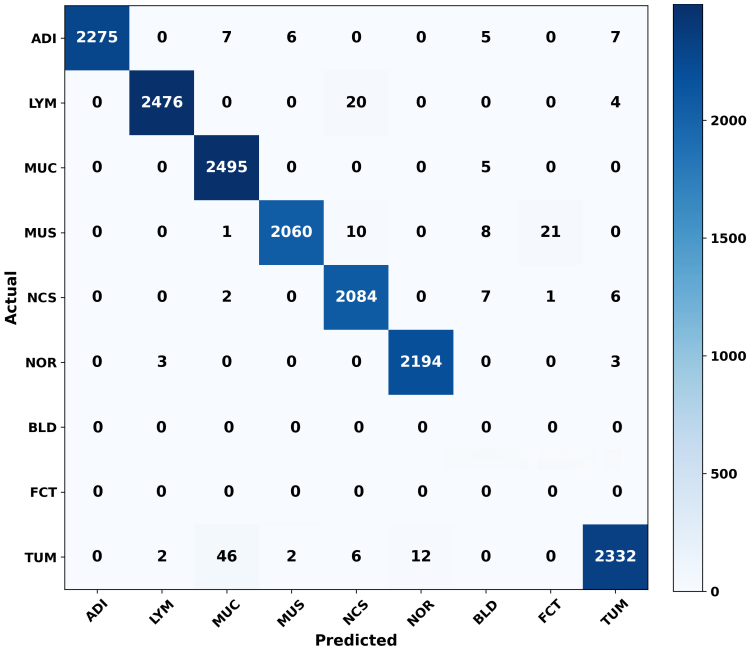
J Confusion matrices for the best-performing models (trained on NCT, HMU, and STARC-9) on CURATED-TCGA-CRC-HE-VAL-20K for seven common tissue types.



(a) NCT



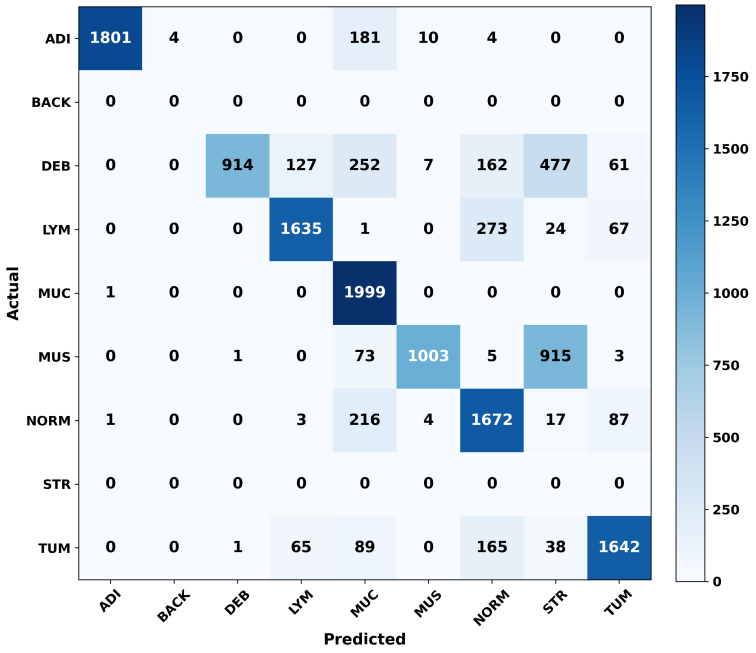
(b) HMU



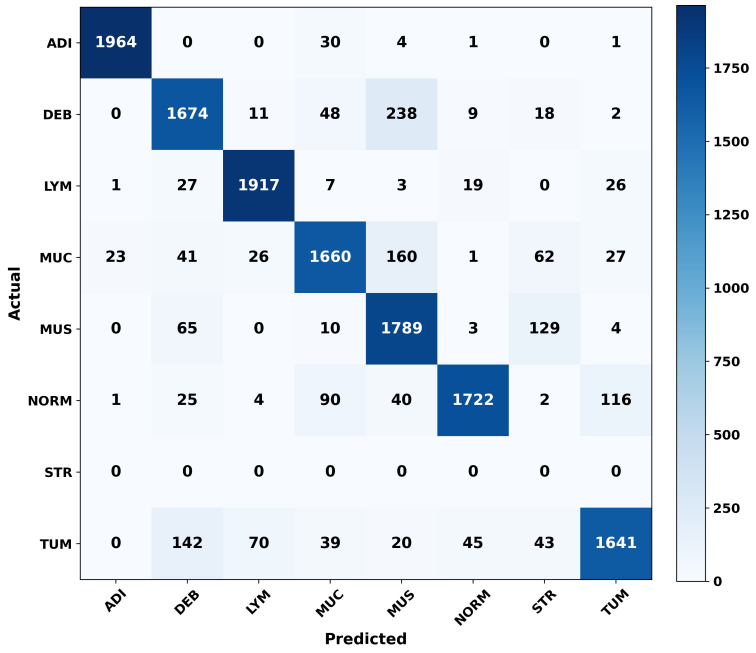
(c) STARC-9

Figure 11: Confusion matrices for the best-performing models on CURATED-TCGA-CRC-HE-VAL-20K for seven common tissue types.

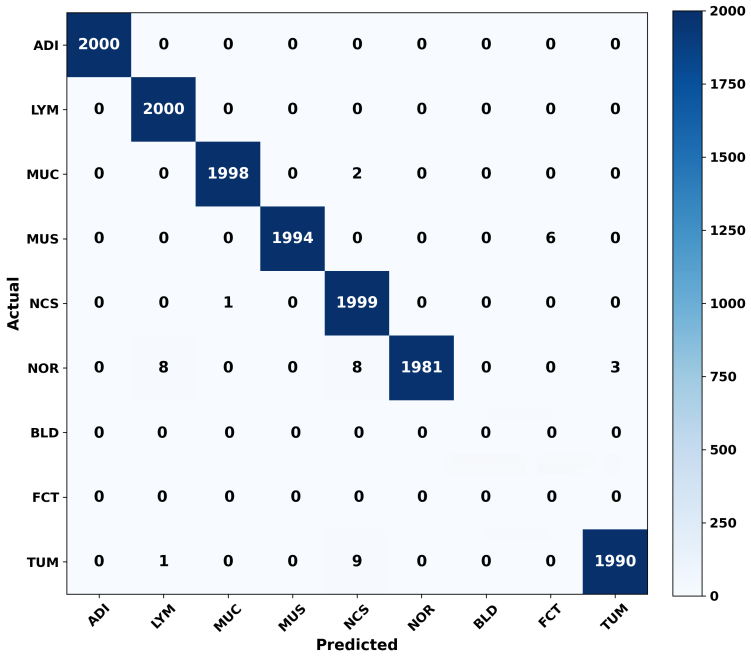
K Confusion matrices for the best-performing models (trained on NCT, HMU, and STARC-9) on STANFORD-CRC-HE-VAL-SMALL for seven common tissue types.



(a) NCT



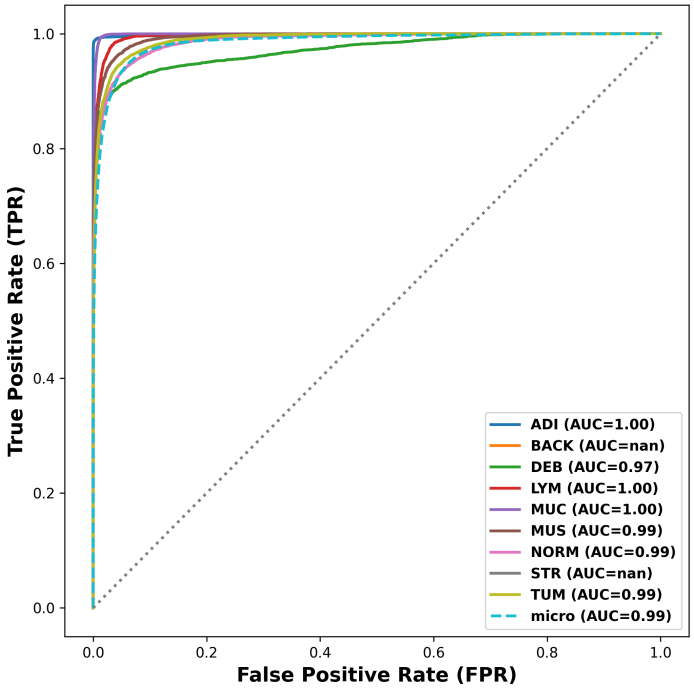
(b) HMU



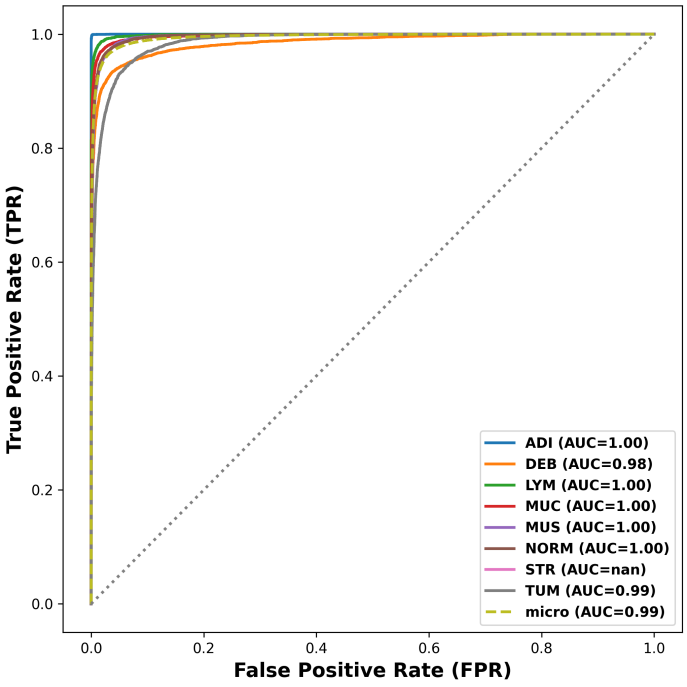
(c) STARC-9

Figure 12: Confusion matrices for the best-performing models on STANFORD-CRC-HE-VAL-SMALL for seven common tissue types.

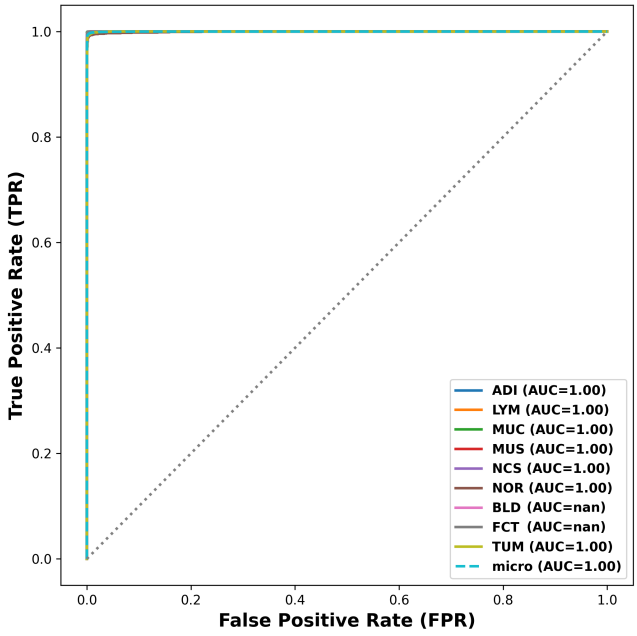
L ROC curves for the best-performing models (trained on NCT, HMU, and STARC-9) on STANFORD-CRC-HE-VAL-LARGE for seven common tissue types.



(a) NCT



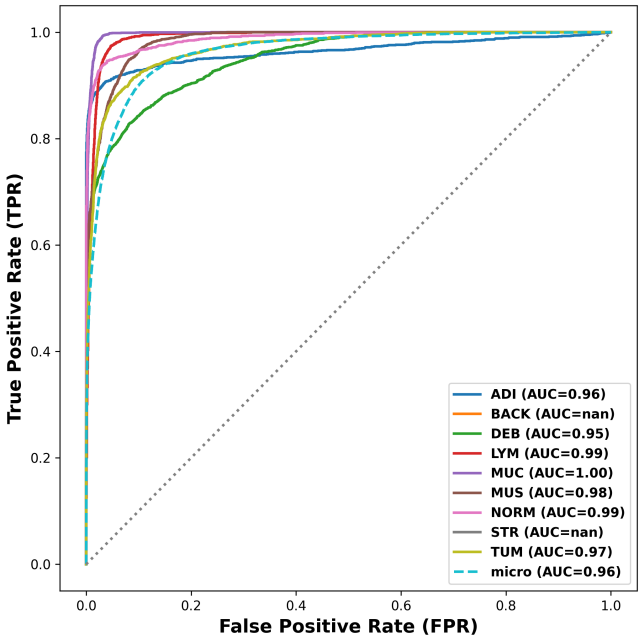
(b) HMU



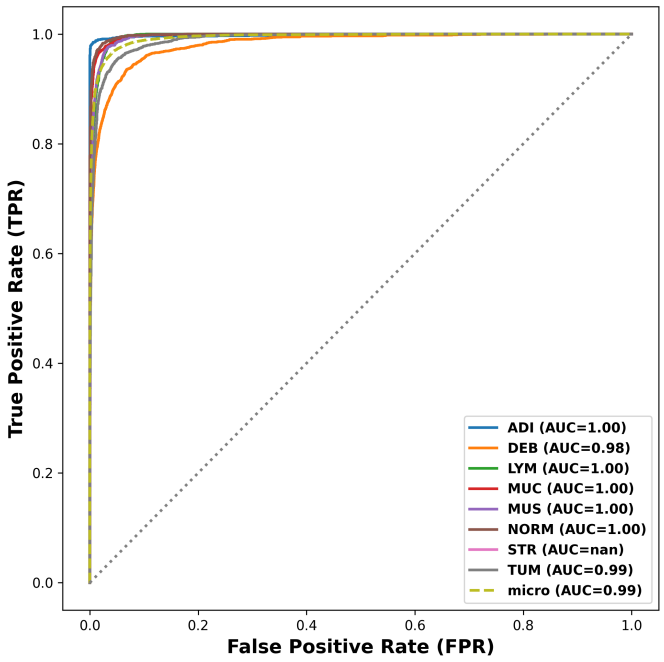
(c) STARC-9

Figure 13: ROC curves for the best-performing models on STANFORD-CRC-HE-VAL-LARGE for seven common tissue types.

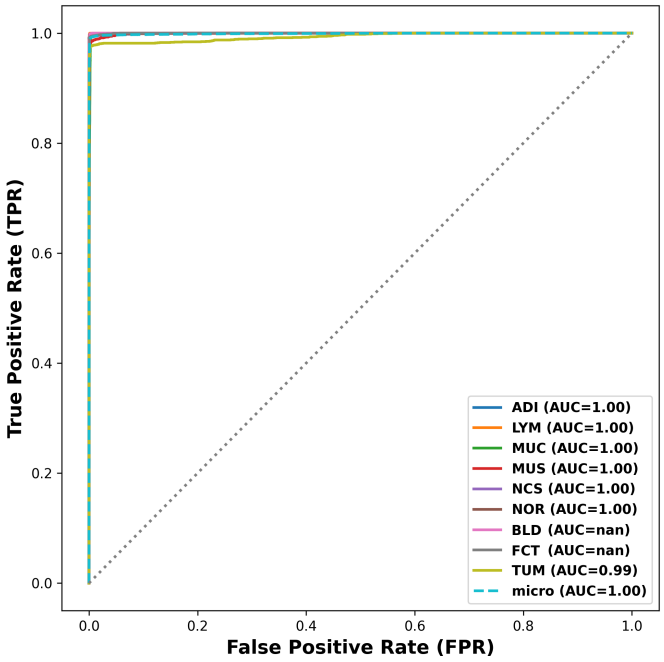
M ROC curves for the best-performing models (trained on NCT, HMU, and STARC-9) on CURATED-TCGA-CRC-HE-VAL-20K for seven common tissue types.



(a) NCT



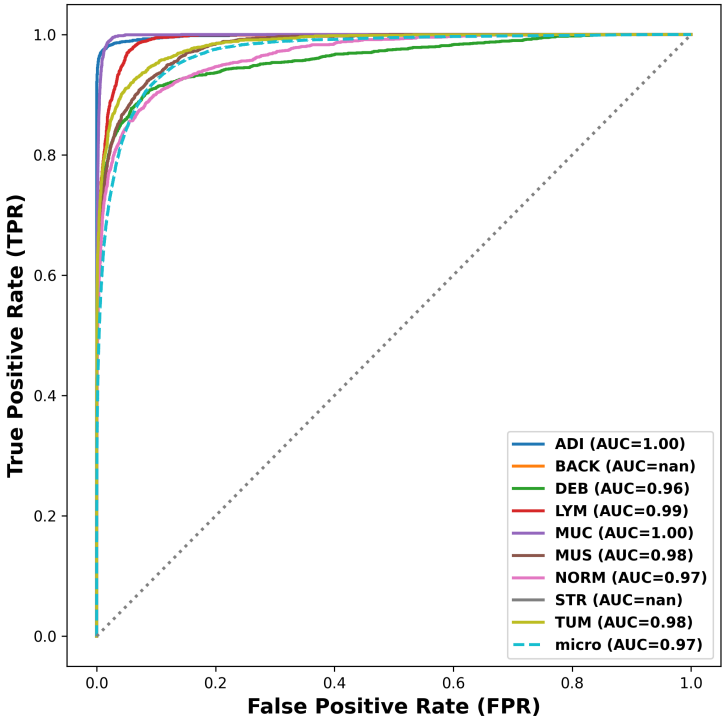
(b) HMU



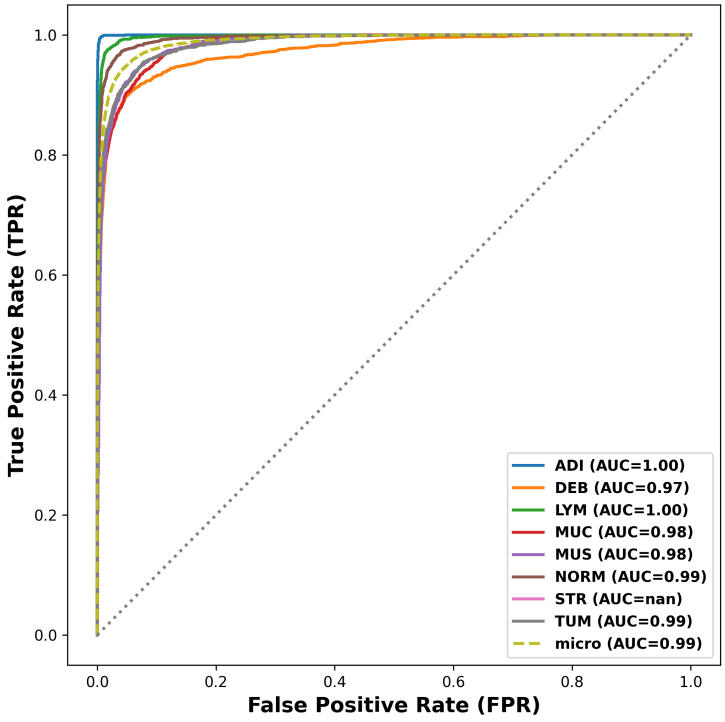
(c) STARC-9

Figure 14: ROC curves for the best-performing models on CURATED-TCGA-CRC-HE-VAL-20K for seven common tissue types.

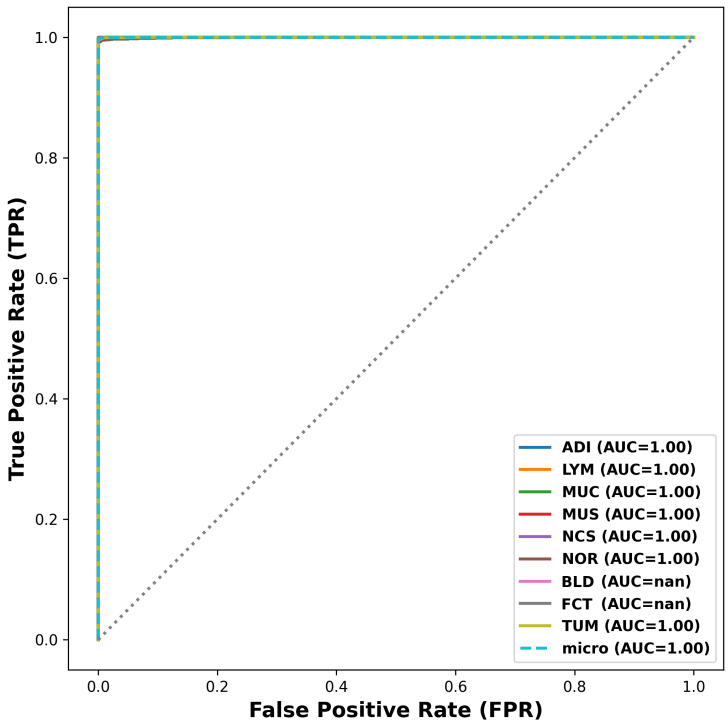
N ROC curves for the best-performing models (trained on NCT, HMU, and STARC-9) on STANFORD-CRC-HE-VAL-SMALL for seven common tissue types.



(a) NCT



(b) HMU



(c) STARC-9

Figure 15: ROC curves for the best-performing models on STANFORD-CRC-HE-VAL-SMALL for seven common tissue types.

O Tumor segmentation within 2048x2048 regions from a WSI from the STANFORD-CRC-HE-VAL-LARGE dataset.

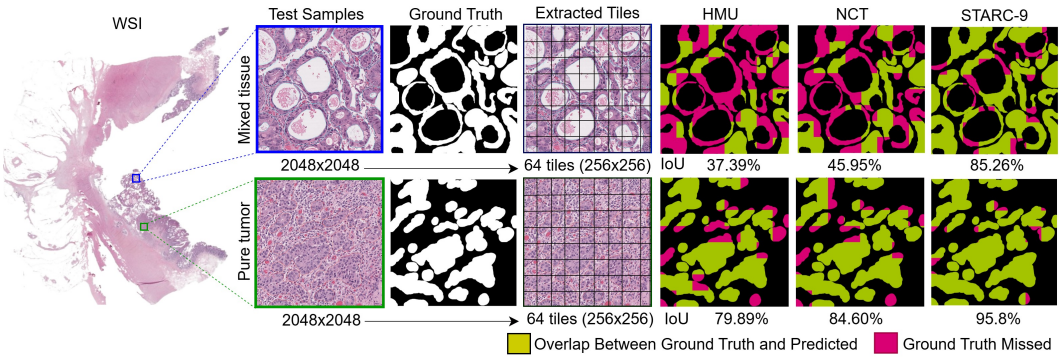
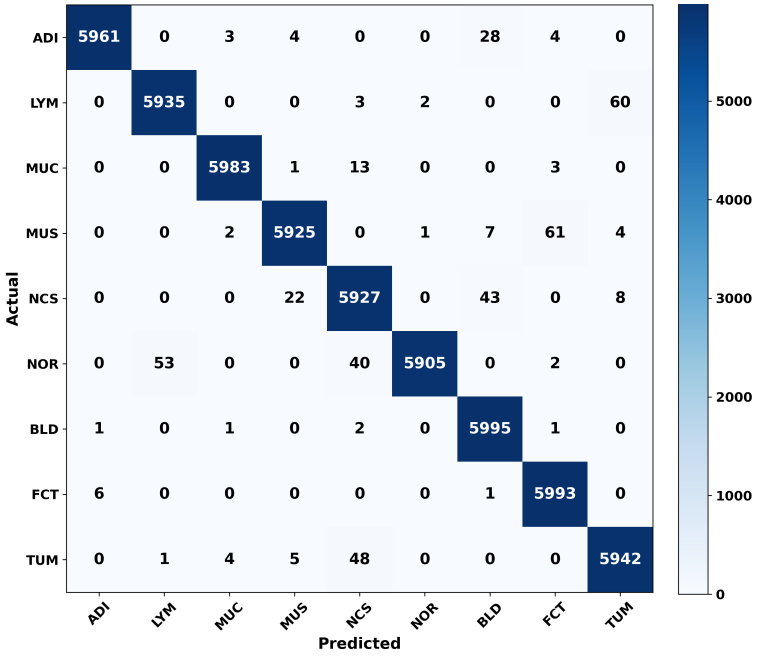
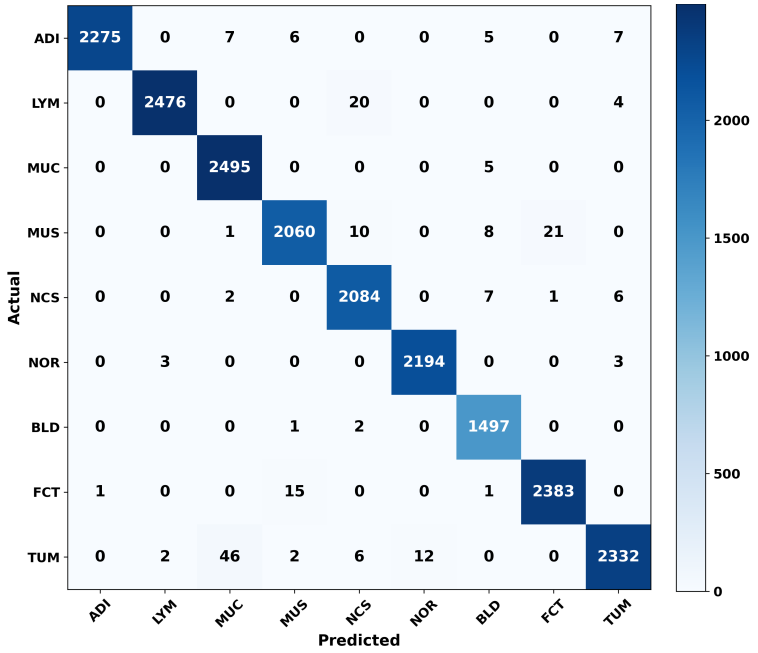


Figure 16: Tumor segmentation within 2048x2048 regions from a WSI from the STANFORD-CRC-HE-VAL-LARGE dataset using tile-level classifiers trained on HMU, NCT, and STARC-9.

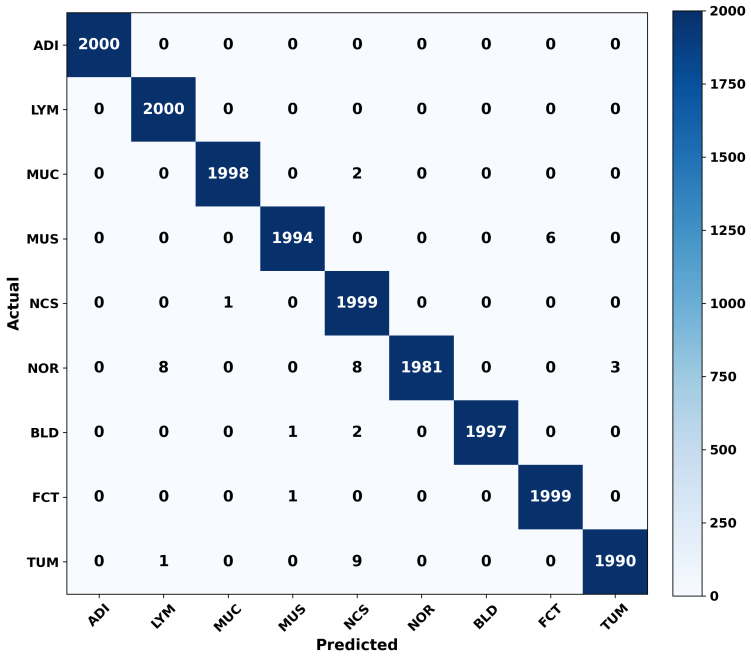
P Confusion matrices for the best-performing model trained on STARC-9 and run on the validation datasets for all nine tissue types.



(a) STANFORD-CRC-HE-VAL-LARGE



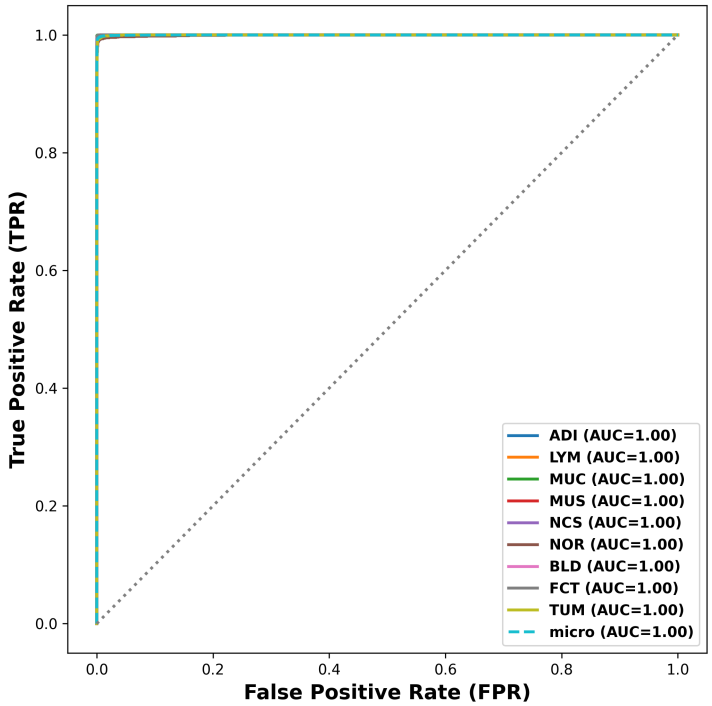
(b) CURATED-TCGA-CRC-HE-VAL-20K



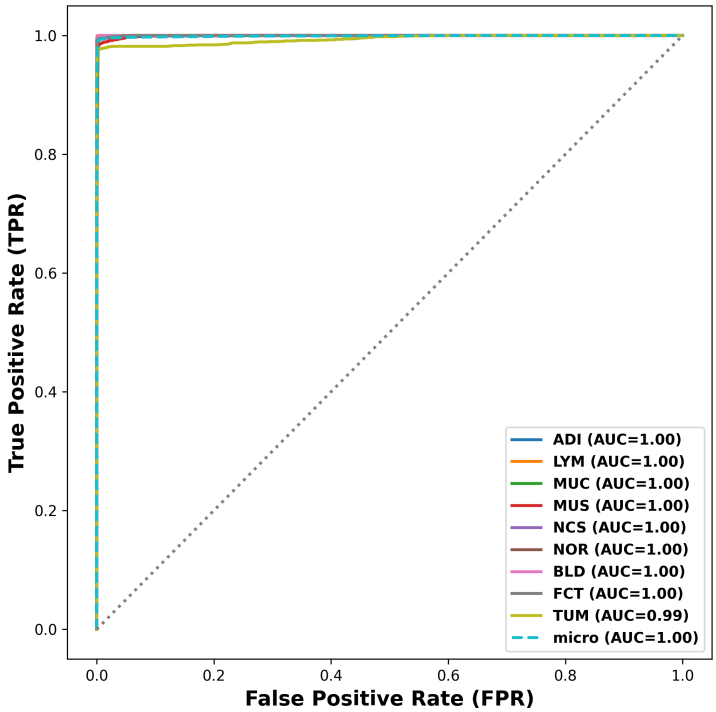
(c) STANFORD-CRC-HE-VAL-SMALL

Figure 17: Confusion matrices for the best-performing model (trained on STARC-9) on (a) STANFORD-CRC-HE-VAL-LARGE, (b) CURATED-TCGA-CRC-HE-VAL-20K, and (c) STANFORD-CRC-HE-VAL-SMALL.

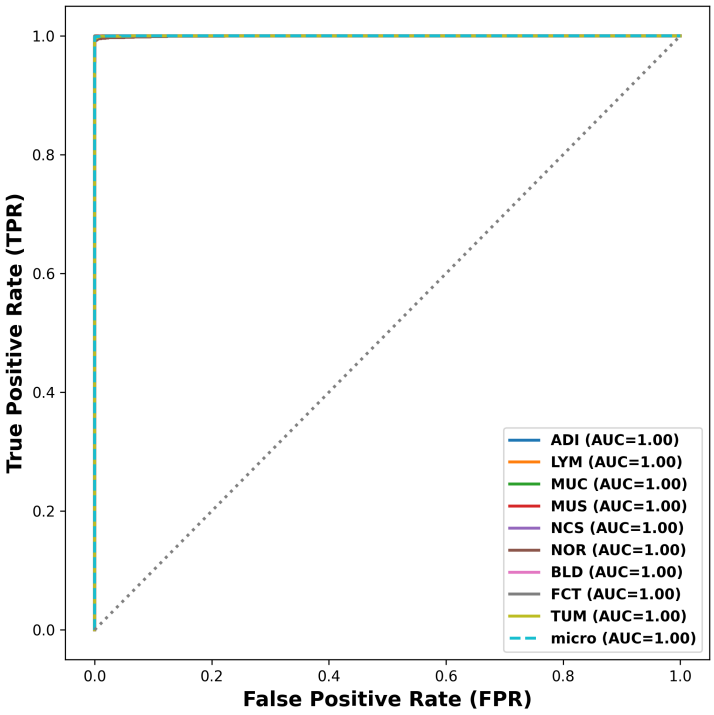
Q ROC curves for the best-performing model trained on STARC-9 and run on the validation datasets for all nine tissue types.



(a) STANFORD-CRC-HE-VAL-LARGE



(b) CURATED-TCGA-CRC-HE-VAL-20K



(c) STANFORD-CRC-HE-VAL-SMALL

Figure 18: Confusion matrices for the best-performing model (trained on STARC-9) on (a) STANFORD-CRC-HE-VAL-LARGE, (b) CURATED-TCGA-CRC-HE-VAL-20K, and (c) STANFORD-CRC-HE-VAL-SMALL.