Chapter 5

Predicting Recurrence Risk from Whole-Slide Images in Early-Stage Breast Cancer

5.1 Introduction

This chapter describes a computational pipeline for assessing recurrence risk in early-stage breast cancer, building on pathology analysis tools introduced in the previous chapter. Early-stage breast cancers are now among the most commonly diagnosed cancers worldwide due in large part to growing investment in cancer screening programs. Many of these patients benefit from favorable prognoses, often receiving therapies specifically tailored to the biological properties of their disease. Patients with hormone receptor (HR)-positive/human epidermal growth factor receptor 2 (HER2)-negative tumors, for example, are a common subgroup shown to benefit substantially from post-surgical endocrine therapy. However, distant recurrence remains a significant concern for many patients, and many receive adjuvant chemotherapy to further minimize the risk of recurrence. In these cases, patients and clinicians must carefully weigh survival benefits against potential cytotoxic side effects.

The clinically validated Oncotype DX™ (ODX) breast cancer assay has become a popular genetic tool that helps clinicians make informed decisions for chemotherapy. This 21-gene assay stratifies patients into low-, intermediate-, and high-risk groups, with higher risk scores correlating with greater benefit from adjuvant chemotherapy. Unfortunately, laboratory-grade genetic testing remains costly and is often delayed or unavailable in low-resource areas, leading many researchers to seek more cost-efficient predictive models. For example, the recently developed Magee Equations™ have successfully predicted ODX risk groups using routinely available clinicopathologic variables.

In this chapter, we investigate the utility of predicting recurrence risk from tissue morphology alone, using deep learning to establish a link between tissue morphology and gene-based risk scores. A trained image encoder, as described in Chapter 4, was used to compress digitized whole-slide images (WSIs) by transforming small image tiles into representative tokens. These tokens were then analyzed by a transformer model to produce risk scores from zero to one. Considering the absence of a large, curated dataset of breast cancer slides with associated ODX scores, we used slides from The Cancer Genome Atlas (TCGA) and their associated gene expression data. Approximate risk scores (RS) were calculated using the 21-gene formula published in the development of the ODX assay. Alongside the predicted recurrence score, several tools were developed to visualize regions of the slide that were most closely associated with high risk. The accuracy of our model was assessed in a separate test set, and image tiles from high-risk regions were collected and clustered together to produce morphological groups of interest. In doing so, we set the stage for deep learning to act not only as an alternative to expensive genetic testing but also to generate novel histopathological hypotheses that can bridge the gap in predictive capacity between histological analysis and genetic testing.

5.2 Background

5.2.1 Assessing Recurrence Risk in Early-Stage Breast Cancers

As healthcare systems worldwide place greater emphasis on regular cancer screening, they must adapt to the growing subset of patients with early-stage disease. In the United States alone, the introduction of mammography screening has doubled the diagnosis rate of early-stage breast cancer, and this trend continues to progress as early detection programs are implemented globally [65]–[67]. While early-stage cases tend to have low mortality rates, careful assessment of the patient's tumor characteristics is required to provide the most effective care while avoiding overtreatment [68], [69]. Breast cancer can be classified according to the expression of key surface proteins: estrogen (ER), progesterone (PR), and the human epidermal growth factor receptor 2 (HER2). The most common subtype is the hormone-receptor-positive, HER2-negative (HR+/HER2-) group, which comprises over 70% of early-stage breast cancer cases [70]. Following surgical treatment, patients in this group typically benefit from targeted therapies that inhibit ER and PR signaling and remain disease-free for many years. However, within this subtype, the risk of recurrence is case-dependent, and care teams must carefully consider several clinical and biological factors to determine who might benefit from more aggressive treatments like chemotherapy [71].

Notably, existing clinicopathological parameters alone have poor prognostic capacity due in large part to inter-observer variability and differences in scoring methods [72], [73]. For example, ER and PR positive status is often determined by immunohistochemistry, requiring manual counting of positively stained cells under a microscope. Over time, this difficulty in assessing risk has resulted in over-treatment and psychological distress for patients who often overestimate the severity of disease [74]. To overcome these prognostic limitations, researchers have developed several tools for assessing risk in early-stage breast cancer. The ODX assay has gained popularity for its ability to predict 10-year recurrence risk for HR+/HER2- patients [75], [76]. The assay produces a risk score (RS) from 1 to 100, split into 3 categories: low risk (0-10), intermediate risk (11-25), and high risk (≥ 26). Large-scale studies have determined that patients in the low-risk group receive little benefit from adjuvant chemotherapy. In contrast, high-risk patients benefit substantially from systemic treatment [76], [77].

5.2.2 Deep learning for cost-effective risk stratification

Though the ODX assay has received widespread acceptance, researchers have long been interested in developing more cost-effective predictive models [78]–[81]. Indeed, ODX assays cost up to USD 4000, limiting their adoption in low-resource settings. The Magee Equations™, for example, use semi-quantitative values for four common immunohistochemistry stains (ER, PR, HER2, and Ki-67) along with mitotic activity scores to reliably generate risk stratifications that match ODX recurrence scores [82]. Ki-67, a cell proliferation marker, and mitotic counts, which measure the number of cells undergoing division, add critical information about tumor growth speed. This complements HR and HER2 status for a more comprehensive risk assessment, offering insights into tumor aggressivity. However, despite established protocol standards, interpreting immunohistochemical scores like ER or HER2 highly depends on pathologist expertise and staining quality. Similarly, assessing mitotic activity requires manual counting of rare mitosis events, which differ in morphology according to the phase of mitosis and can often be confused with apoptosis events.

Researchers in computational pathology have long sought to tackle observer-subjectivity by leveraging deep learning algorithms to produce more thorough and consistent results in whole slide image analysis[83]–[86]. In breast cancer, specifically, separate deep learning studies have been successful in predicting each of the Magee Equation™ metrics from digitized histology slides alone [87]–[91]. This research demonstrates that specific cell and tissue structure patterns, referred to as "morphologic signatures," are associated with widely used risk assessment metrics, supporting the hypothesis that a deep learning model might predict recurrence risk directly from digitized histology slides.

In this chapter, we describe a deep learning approach that can serve as a cost-effective alternative to expensive genetic testing while avoiding the limitations of manually calculated histopathological metrics. We leverage Howard et al.'s "research-based" Oncotype DX™ dataset as our training set. This dataset assigns risk scores to digitized pathology slides from the Cancer Genome Atlas (TCGA) using published Oncotype DX™ formulas and TCGA gene expression data [92]. To perform whole-slide analysis, we first divide the foreground region of the slides into arrays of smaller tiles and compress each tile into a representative vector using the trained encoder outlined in Chapter 4. This preserves the spatial relationships between tiles while retaining only the salient features of the tile. The representative vectors created for each slide were then analyzed using a vision transformer architecture, which excels in capturing long-distance dependencies within an image due to its Multi-Head Self-Attention (MHSA) mechanism [93].

Vision Transformer (ViT) models build on the groundbreaking success of transformer models, which revolutionized natural language processing by demonstrating the effectiveness of self-attention mechanisms in generating coherent, context-aware text. To adapt the transformer architecture to image analysis, ViTs split an image into a sequence of smaller image patches, or tokens, reading each token similar to words in a sentence. The self-attention mechanisms of the transformer architecture enable ViTs to evaluate and weigh the importance and relationship of each image patch relative to the rest. This allows the model to dynamically focus on different parts of the entire image at each layer, facilitating a detailed understanding of the image's content by considering both local details and global context. In contrast to convolutional neural networks (CNN), which rely on hierarchical convolutional layers to progressively expand receptive fields, ViTs consider interactions between all regions of an image at each step, as illustrated in Figure 5.1. This crucial difference in analysis has led to significant improvements in image analysis across domains [94], [95]. Vision transformers have demonstrated success in many histopathology applications, where key morphological patterns often emerge across long-range distances [96]–[99]. Once trained, models like vision transformers offer a scalable solution with significantly lower operational costs than laboratory testing, highlighting the potential for deep learning to improve access to vital risk assessment tools.

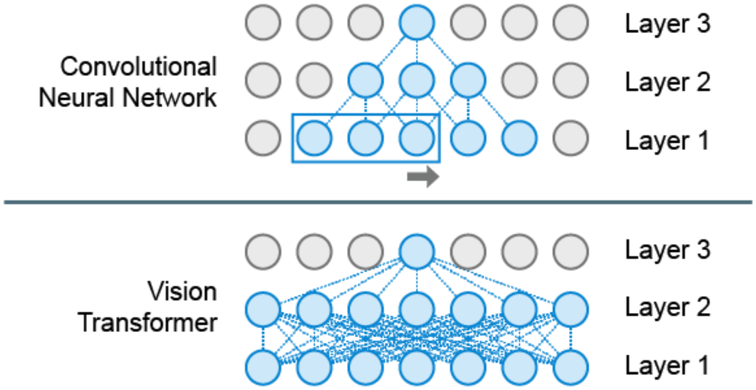

Figure 5.1: **Convolutional Neural Networks vs. Vision Transformers for Image Analysis.**(Above) Convolutional neural networks (CNN) use a moving convolutional filter to generate new layers in the model. In deeper layers, the nested convolutions allow each neuron to access a greater portion of the input layer. (Below) In contrast, vision transformers (ViT) consider all regions of the input layer to generate each neuron in the next. This design allows ViTs to consider long-range relationships more efficiently.

5.2.3 Model Interpretability: Mapping Recurrence Risk to Tissue Morphology

As demonstrated in previous chapters, to effectively integrate AI tools into clinical and diagnostic workflows, models must be able to provide interpretable or verifiable results. For risk assessment using whole-slide images, it is important for both researchers and clinicians to identify specific regions on the slide that lead the model to produce higher risk scores. By reviewing these areas, clinicians can integrate AI insights with their domain expertise and additional clinical data not considered by the model. Whereas in a research context, these regions can provide valuable insights into the underlying biological mechanisms driving aggressive disease or response to therapy.

Producing explainable results from vision transformers and other deep learning models is an active area of research [100]–[103]. Specific to ViTs, one approach lies in exploring the model's self-attention layers. These layers compare all tokens against each other to compute attention scores, allowing ViTs to focus on different regions as the input tokens are transformed. Attention-based explainability techniques synthesize these attention weights across model layers to reveal which input sections are most influential in decision-making [104]. However, the set of tokens with high attention may be too broad to provide any insight, or it may include tokens that are important for "ruling out" decisions. Thus, this technique alone may be unreliable for understanding which slide regions lead to elevated risk scores.

Perturbation techniques do not depend on model architecture and involve intentionally altering parts of the input data to understand how they impact model outputs. Through randomly generated occlusions or noise injections, perturbation approaches to model explainability offer more direct insight into which regions contribute most heavily to model outputs. These techniques provide the benefit of generating input-resolution maps but are computationally intensive and must be carefully tuned to each model's sensitivity to noise. Perturbations may be too aggressive or subtle, leading to over- or under-estimations of the importance of certain regions.

In the work described below, we leverage both techniques (self-attention and perturbation) to build a more robust map of slide regions associated with high-risk scores. Furthermore, design choices in our analysis pipeline allow our model to generate coarse-grained risk prediction maps alongside slide-level predictions automatically. We split slides into large overlapping regions, each processed individually using our transformer architecture and given individual risk scores. During training, we leverage specialized loss functions from the multiple instance learning (MIL) field to ensure that local-level predictions are compatible with globally assigned risk scores. With these three interpretability metrics, we generate risk heatmaps with greater confidence and are better equipped to provide clinically interpretable risk assessments.

5.3 Materials and Methods

5.3.1 The Cancer Genome Atlas Pseudo Recurrence Score

For our dataset, we used the published "research-based" risk scores generated by Howard et al. [92]. We selected HR+/HER2- patients for training and testing. Due to data limitations within the TCGA database, menopausal status was not considered in our dataset curation. A total of 1,039 patients, with 1,099 associated slides, were used for training and testing. Only patients with one or no lymph node metastases were selected. Further patient demographic and pathologic parameters are described in Table 5.1. A test set of 100 patients was allocated randomly, leaving 939 patients for training and validation. A validation split of 10% was used during training.

To generate ground-truth values model training, research-based risk scores (RS) were mapped to a range from 0.0 to 1.0 according to:

$$r_i = f(RS_i) = \begin{cases} 0.5\left(\frac{RS_i-\alpha}{\theta-\alpha}\right)^{\tau} & \text{for } RS_i \leq \theta \\ 1-0.5\left(1-\frac{RS_i-\theta}{\beta-\theta}\right)^{\tau} & \text{for } RS_i > \theta \end{cases}$$

Where α and β are the minimum and maximum RS values, respectively, θ is the cutoff for RS categorized as high-risk, and τ=1.5 is the parameter used to control the separation between low-risk and high-risk groups. Higher values of τ push the low-risk and high-risk scores further apart from each other.

Table 5.1: Demographic statistics for training dataset for research-based ODX risk score prediction

Category	**Subcategory**	**Count**	**%**
Age	Median (min-max)	59.0 (26 - 90)	
Race	White	774	69.50%
	Black	167	15.60%
	Asian	60	5.80%
	Native American	1	0.10%
	N/A	94	9.00%
Risk Group	Low	737	70.93%
	High	302	29.07%
Histology Broad Groups	Ductal and lobular neoplasms	245	23.58%
	Carcinoma (not specified)	746	71.80%
	Other	48	4.62%
Histological Grade	I	229	22.04%
	II	428	41.19%
	III	380	36.57%
	Unknown	2	0.19%
Tumor Grade	T1	263	25.31%
	T2	606	58.33%
	T3/4	168	16.17%
	Unknown	2	0.19%
Tumor Size	≤1.0 cm	57	5.49%
	1.0–2.0 cm	211	20.31%
	2.0–4.0 cm	607	58.42%
	≥4.1 cm	164	15.78%
ER	positive	766	73.72%
	negative	226	21.75%
	Unknown	47	4.52%
PR	positive	662	63.72%
	negative	329	31.67%
	Unknown	48	4.62%

5.3.2 Slide Pre-Processing

As shown in Figure 5.2, connected tissue regions from each set of patient slides were isolated and converted to vector embeddings. We denote each patient as p_i, with an associated set of whole-slide images, s_i. From s_i, contiguous tissue regions were isolated from the foreground and split into non-overlapping square tiles of size (256 × 256) or (512 × 512) pixels for 20X and 40X magnifications, respectively. Larger tiles were resized to (256 × 256). The pre-trained encoder described in Chapter 4 produced 2048-length embeddings for each tile. The array of embeddings representing each contiguous tissue region was then split into uniformly sized arrays of dimension $(S \times S \times D)$ where $S = 32$ and $D = 2048$ is our embedding dimension, using overlapping crops or padding with blank space as necessary. Thus, each patient p_i was represented by an array $x_i \in \mathbb{R}^{n_i \times S \times S \times D}$, where n_i is the number of $(S \times S \times D)$ regions.

A set of 1000 background images from random slides was collected, encoded, and averaged to produce the background (BG) tile embedding used for padding throughout this work. All embeddings were zero-mean and unit-variance normalized. Whole-slide images were read with the OpenSlide library for Python 3.8 [57]. The lowest resolution image for each file was used to generate a foreground mask using the Python HistomicsTK library [105].

5.3.3 Down-scale Vision Transformer for Whole-Slide Analysis

As described in section 5.3.2, each patient p_i is represented by a collection of embedding arrays $x_i \in \mathbb{R}^{n_i \times S \times S \times D}$. This set of arrays is processed by a vision transformer, as shown in Figure 5.3. A small neural network $\phi(x)$, consisting of a dense layer with zero-bias followed by GELU activation, was applied to fix embedding dimensionality at d=128 [106]. We employ 2-dimensional sinusoidal positional encodings to imbue our array with positional information [107]. Following this step, the array is processed by 13 multi-scale transformer blocks, as introduced by Li et al. [108]. Table 5.2 outlines transformer block parameters. In the final step, our transformed token arrays are consolidated by concatenating average and maximum pool layers.

The input x_i is thus transformed as follows:

$$\begin{aligned} x_i' &= \phi(x_i + PE(x_i)) \in \mathbb{R}^{n_i \times S \times S \times d} \\ x_i'' &= T(x_i') \in \mathbb{R}^{n_i \times \frac{S}{2} \times \frac{S}{2} \times d} \\ x_i''' &= [\mathrm{AvgPool}(x_i''), \mathrm{MaxPool}(x_i'')] \in \mathbb{R}^{n_i \times 2d} \end{aligned}$$

where $T(x) = T_{12}(\cdots T_0(x))$ represents the 13 transformer-block function, $\phi(\cdot)$ represents the dimensionality reduction neural network, and $PE(\cdot)$ represents the sinusoidal positional encoding function.

Our consolidated region-level embeddings, denoted as x_i''' or $\{x_{i,j}'''\}_{j=1}^{n_i}$, are transformed into a global risk prediction r_i using attention-weighted averaging. A neural network consisting of two dense layers is applied to create attention weights for each region, and the attention weights are used to create a final patient-level representation. The patient-level representation, as well as the region-level embeddings themselves, are transformed into risk predictions using a final dense layer. The risk prediction steps are as follows:

$$z_i = \frac{\sum_{j=1}^{n_i} a_j \cdot x''_{i,j}}{n_i} \in \mathbb{R}^{2d}, \quad \text{where } a_j = \sigma(A(x'''_{i,j}))$$
$$\tilde{r}_i = R(z_i) \in \mathbb{R}$$
$$\tilde{r}_{i,j} = R(x'''_{i,j}) \in \mathbb{R}$$

where $A(\cdot)$ is the neural network that generates attention weights, σ is the SoftMax function, n_i is the number of tissue regions associated with patient p_i, and $R(\cdot)$ is a dense layer for class prediction.

Table 5.2: Transformer Block Parameters

Block #	Type	Kernel	Stride	Input Size	Block DropPath	Attn Heads	Attn Dropout	MLP Units	MLP Dropout
0	Downscale	(3, 3)	(2, 2)	(S, S, d)	0.0	1	0.0	256	0.0
1	Standard	-	-	(S/2, S/2, d)	0.033	2	0.1	256	0.4
2	Standard	-	-	(S/2, S/2, d)	0.066	2	0.1	256	0.4
3	Standard	-	-	(S/2, S/2, d)	0.1	2	0.1	256	0.4
4	Standard	-	-	(S/2, S/2, d)	0.133	2	0.1	256	0.4
5	Standard	-	-	(S/2, S/2, d)	0.166	2	0.1	256	0.4
6	Standard	-	-	(S/2, S/2, d)	0.2	2	0.1	256	0.4
7	Standard	-	-	(S/2, S/2, d)	0.233	2	0.1	256	0.4
8	Standard	-	-	(S/2, S/2, d)	0.266	2	0.1	256	0.4
9	Standard	-	-	(S/2, S/2, d)	0.3	2	0.1	256	0.4
10	Standard	-	-	(S/2, S/2, d)	0.333	2	0.1	256	0.4
11	Standard	-	-	(S/2, S/2, d)	0.366	2	0.1	256	0.4
12	Standard	-	-	(S/2, S/2, d)	0.4	2	0.1	256	0.4

5.3.4 Model Training

We trained our custom transformer model one patient at a time, with the first dimension of x_i acting as the batch dimension for simplicity. The loss function used during training combined three major components: i) the global prediction loss, ii) the regional prediction loss, and iii) the regularization loss.

The global prediction loss is calculated using standard categorical cross entropy as follows:

$$L_{\text{global}} = BCE(\tilde{r}_i, r_i),$$
$$BCE(\tilde{r}_i, r_i) = -(r_i \log \tilde{r}_i + (1 - r_i) \log(1 - \tilde{r}_i))$$

where r_i and $\widetilde{r}_i$ are the ground truth and predicted risk scores for patient p_i, respectively.

To calculate regional prediction loss, it was assumed that high-risk patients must have some minimum fraction of the tissue regions with high-risk scores. Thus, the regional prediction loss is calculated by:

$$L_{\text{regional}} = \frac{1}{|K_i|} \cdot \sum_{j \in K_i} BCE(\tilde{r}_{i,j}, r_i)$$

where the set K_i represents the top 10th of regions with the highest scores. The L1 and L2 regularization components were calculated as follows:

$$L_{reg} = L_1\left(\theta_\phi,\ \alpha\right) + L_2\left(\theta_T,\ \beta\right) + L_2\left(\theta_A,\ \beta\right) + L_2\left(\theta_R,\ \beta\right)$$

where $\theta_\phi, \theta_T, \theta_A, \theta_R$ denote the model parameters for the $\phi, T, A,$ and R networks, respectively, and $\alpha = 5 \times 10^{-4}, \beta = 1 \times 10^{-5}$ are the L1 and L2 regularization coefficients, respectively. The three components were combined as follows:

$$L(r_i, \tilde{r}_i, \tilde{r}_{i,j}) = (1-\lambda) \cdot L_{\text{global}} + \lambda \cdot L_{\text{regional}} + L_{\text{reg}}$$

$$\omega(r) = \begin{cases} 2.17 & \text{if } r \geq 0.5 \\ 0.65 & \text{if } r < 0.5 \end{cases}$$

where $\lambda = 2 \times 10^{-1}$ is the coefficient used to weigh global against regional loss, and ω is our class weight function, used to offset class imbalance between low- and high-risk scores.

The model was implemented using the Tensorflow machine learning library, Version 2.13.0, on Amazon Web Services (AWS) Elastic Compute Cloud (EC2) g5.16xlarge instances [59]. Our model was trained for 100,000 steps using the Adam optimizer with a learning rate of 1×10^{-3}.

5.3.5 Augmentation Techniques

The input array $x_i \in \mathbb{R}^{n_i \times S \times S \times D}$ was augmented by random horizontal and vertical flipping, and 90° rotations. As described in Figure 5.4, a second augmentation step was applied to introduce noise to our input. First, the input was randomly split into blocks of size 8 to 16 along the height and width dimensions. Then, up to 100% of each region (along axis 0 of x_i) was selected for augmentation. For the selected blocks in each region, tokens were replaced with background tokens or an average token. The second option replaced block tokens with an average of those in that block. As a final augmentation step, Gaussian noise was introduced to each token, a common augmentation step in image processing. A standard deviation value of 0.1 was used for Gaussian noise.

5.3.6 Visualizing High Risk Regions

Attention maps were created using the attention weights from transformer blocks T_1 to T_{12}, as well as the attention weights assigned to each region, a_j for $j \in \{1, \cdots, n_i\}$ when generating patient-level results (see 5.3.5). Transformer attention weights were max reduced across attention heads, and the joint attention was calculated for each region, as described by Abnar et al. [104]. A filtering step removed all attention below the 10th percentile. Finally, the joint attention matrix for each region was reduced by aggregating the attention received by each token. Thus, each region produced a joint attention matrix $a_j^T \in \mathbb{R}^{\frac{S}{2} \times \frac{S}{2}}$. The region-level attention weights $a_j \in \mathbb{R}$ were added to each a_j^T and mapped onto the original slide. This map was max normalized and filtered to keep the top 15th percentile.

Drop-out analysis was conducted by generating perturbed versions of the input x_i, and recording changes to the model output. As described in 5.3.6, the input was split into blocks and up to 75% of the blocks were chosen to be replaced with BG tokens. The model output from the perturbed input was compared to the original prediction $\widetilde{r}_i$ and the difference was recorded for those perturbed blocks. After $n_{iter} = 100$ iterations, the differences recorded for each token were averaged. Tokens with low $\Delta\widetilde{r}_i$ values indicated regions where occlusion led to lower risk scores and were thus interpreted as high-risk regions. This map was min-max normalized and filtered to keep the top 15th percentile. Region-level predictions, $\widetilde{r}_{i,j}$, were also mapped to the original slide space to generate a coarse risk map. Combined attention maps were created by adding all three maps and filtering to keep the top 0.1%. A Gaussian filter was applied to the combined map to reduce sharp edges for visualization.

5.3.7 Clustering High-Risk Regions

Risk maps were collected from all test set examples correctly predicted to be high-risk. Non-zero regions of the map were collected as image tiles of size 256×256 and encoded using our histology image encoder. The uniform manifold approximation and projection (UMAP) algorithm in the Python Sci-kit Learn library was applied to project these vectors to a lower 64-dimensional space prior to clustering, and the following parameters: `n_neighbors`=10, `min_dist`=0.0 were used [36]. To distinguish clusters, we applied the HDBSCAN algorithm, a hierarchical density-based approach that can detect clusters of arbitrary shape and density [63] `min_cluster_size`=50, `min_samples`=2. A separate pseudo-cluster of benign tissue was created using known benign tissue in the annotated BreAst Cancer Histology images (BACH) dataset [42].

Tumor classification scores for each tile in each cluster were calculated using the trained tumor subtyping model described in Chapter 4.

5.3.8 Tumor Region Annotations

Tumor likelihood maps were made using the trained tumor subtyping model described in Chapter 4. The image tiles in the foreground of the input slide were analyzed by this model, converted to tumor likelihood scores, and mapped to the original slide space. The tumor likelihood was taken by summing the probability of an image belonging a malignant class (ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma). Ground truth annotations were taken from an online dataset published for TCGA slides [109].

5.4 Results and Discussion

5.4.1 Dataset Preparation

In this study, we leverage ‘research-based’ risk scores published by Howard et al. to train a neural network to predict recurrence risk from digitized histology slides [92]. With gene expression data from breast cancer patients in the TCGA database, risk scores were calculated using published OncotypeDx™ formulas and assigned to each patient. Researchers used all breast cancer patients in the TCGA database to generate research-based risk scores

(RBRS), and the 15th percentile among HR+/HER2- patients was used to create a high-risk cutoff, matching the percentile of patients with ODX scores greater than 26. To match the patient subtype indicated for OncotypeDx™ testing, a total of 1,039 HR+/HER2- patients with an associated 1,099 slides were selected for training and testing. With 302 patients, the high-risk group represented 29% of all samples. The patient cohort's demographic and pathological characteristics are presented in Table 5.1.

To train our neural network as a binary classifier, research-based risk scores needed to be mapped to a range between 0 to 1.0. Low-risk patient scores were normalized to a range from 0 to 0.50, and high-risk patient scores were normalized from 0.50 to 1.0 (See Section 5.3.1). A mapping function was used to adjust the separation between low and high scores, using a tunable hyperparameter τ to control the tightness of each group's distribution. In setting this parameter, we found that higher values of τ that pushed scores closer to binary values led to greater risks of overfitting, as the training task became simpler, and performed best with a value $\tau =$ 1.5.

As each patient was associated with a variable number of tissue sections, a pre-processing step was designed to facilitate model training. In a given WSI, we considered spatially separated tissue sections independent from each other, as spatial relationships could not be inferred between sections without physical contact. Contiguous tissue regions were split into smaller tiles and mapped to vector representations using our pre-trained histology image encoder (Chapter 4). To simplify computational processing, the resulting arrays of variable height and width representing each tissue region were split into overlapping regions of uniform size, as shown in Figure 5.2A. In this format, all slide regions could be processed in parallel as a single "batch." We chose the dimension S to maximize the area of tissue that could be considered at once under our computational constraints. The chosen value S = 32 corresponded to a tissue section of approximately 4 mm × 4 mm (Figure 5.2B), larger than many smaller tumor samples in our dataset.

5.4.2 Model Design and Training

Pre-processed tissue regions representing each slide were analyzed using our customized ViT architecture, as outlined in Figure 5.3. A transformer architecture was chosen due to its strength in considering global contexts and long-distance interactions between different image regions. This property is fundamental in macro-scale histological analysis, where finer-grain details matter less than the overall shape, distribution, or frequency of specific structures [97]. For instance, when assessing immune infiltration, we would want a model that focuses less on the exact morphology of each infiltrate cluster but rather on their prevalence across the tissue sample. In our model, we specifically used the multi-scale transformer architecture outlined by Li et al. [108]. This design allowed us to re-introduce some local processing by including depth-wise convolutions in the self-attention step.

In our first transformer block, a (3x3) convolution with stride 2 was used to consolidate our token number 4-fold, further minimizing computational loads. After the remaining 12 transformer blocks, region-level representations were attention-weighted and averaged to produce a final patient-level representation. Both region-level and patient-level representations were processed with the same dense layer to produce region-level and patient-level risk predictions, $\widetilde{r}_{i,\ j}$ and $\widetilde{r}_i$, respectively. In producing region-level predictions, we explored

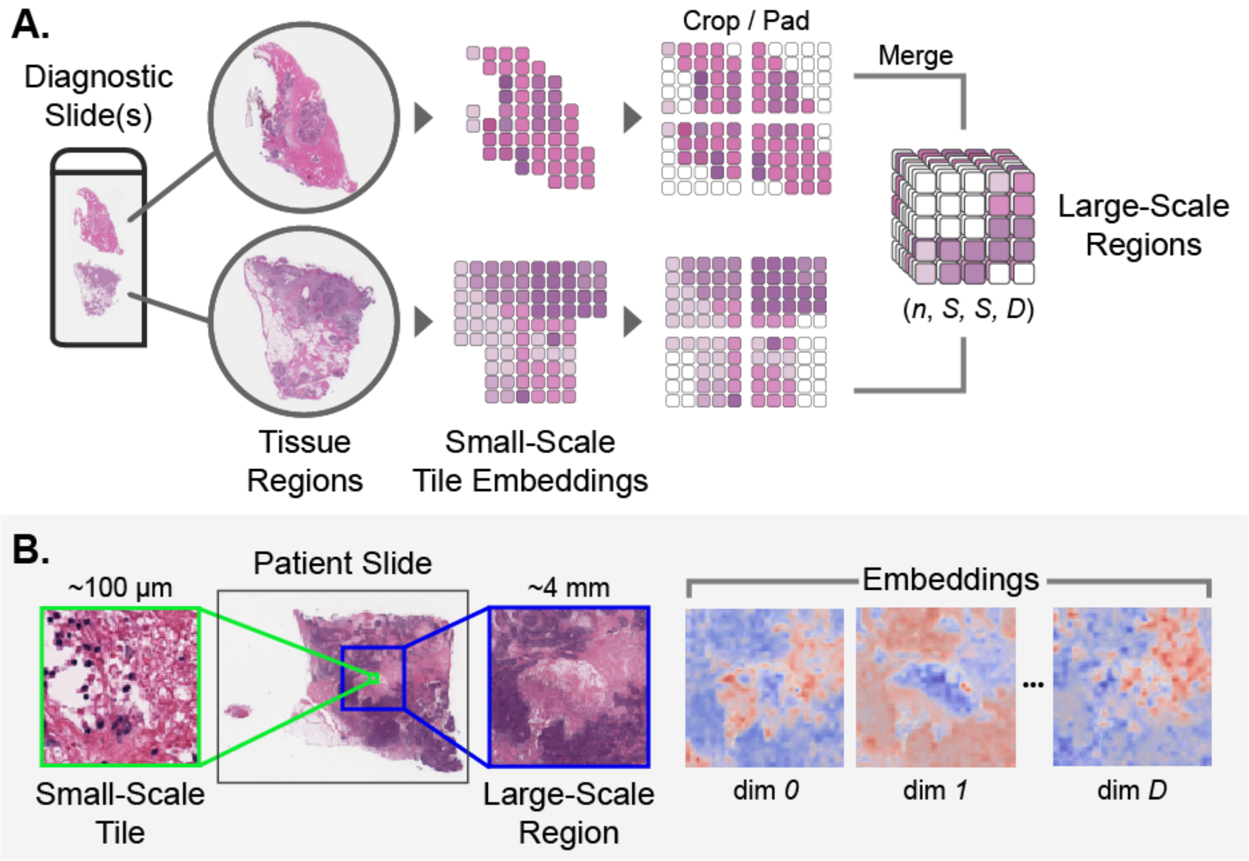

Figure 5.2: **Patient Slide Pre-Processing.**(A) We demonstrate the pre-processing steps for a single-slide example with two contiguous tissue regions. Each region is isolated and split into smaller image tiles that are then embedded using a pre-trained encoder. This creates variably sized arrays of tile embeddings that are then split into uniformly sized regions of size (S x S). Padding is done with a background vector embedding to represent spaces outside the tissue foreground. (B) An example slide demonstrates the scale difference between individual tiles and large-scale tissue regions. On the right, we show the (S x S) heatmaps for each dimension of the embeddings.

our model's capacity to assess risk from a single tissue region, with no global context. Furthermore, region-level predictions would facilitate establishing a connection between specific slide sections and the patient's predicted risk score.

We trained our model for 10,000 steps using a batch size of 1. Our augmentation pipeline was designed to minimize overfitting risks often associated with transformer model. During training, each tissue region representation was augmented with random flips, rotations, and the addition of Gaussian noise. Furthermore, randomly sized blocks were selected from each region and replaced with background or average tokens, as shown in Figure 5.4. This augmentation step was used to simulate either the complete removal of a tissue chunk or the removal of fine details within a chunk due to blurring. This step encouraged our model to leverage a broader, more robust set of features and structural patterns rather than relying on specific local patterns. The loss function employed during training compared the predicted

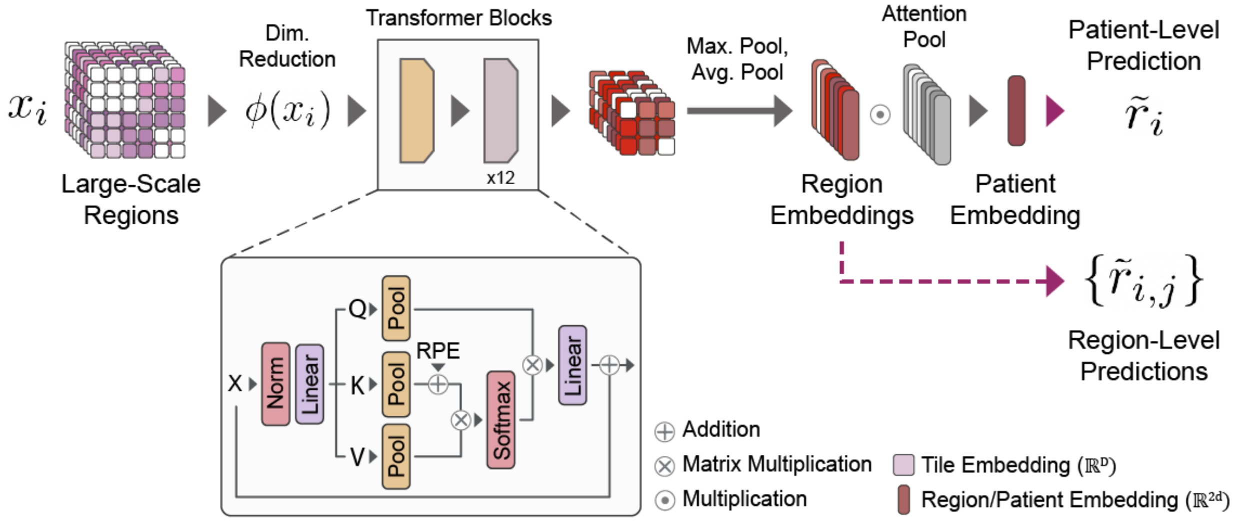

Figure 5.3: **Risk Prediction Model Architecture.**A custom transformer architecture is used to analyze patient data as a collection of large-scale regions and produce region-level and patient-level risk scores. The patient data x_i is first transformed by a phi network to reduce tile embedding dimensionality. These embedding, or tokens, were then processed by 13 transformer blocks, the first of which introduced a convolution pooling step to consolidate local windows of tokens. Region-level embeddings were created by concatenating the results from max. and mean pooling on our transformed tokens. Region embeddings were attention-weighted and averaged to produce a final patient embedding. Both region and patient embeddings were mapped to a final risk score.

patient-level risk score $\widetilde{r}_i$ to the actual risk score r_i. Furthermore, the top 10th of regions with the highest regional predictions, $\widetilde{r}_{i,\ j}$, were compared to r_i, under the assumption that a minimum fraction of regions must be considered high-risk if a patient belongs to the high-risk group.

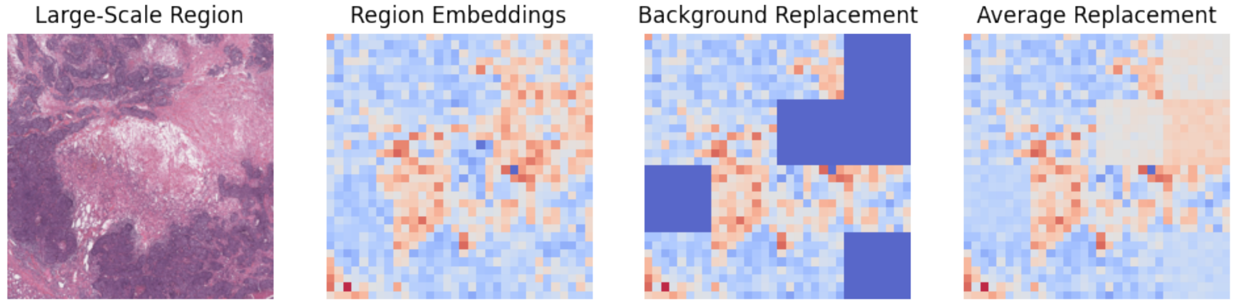

Figure 5.4: **Input Augmentation for Model Training.**A single tissue region is shown along with the heatmap for the embeddings at dimension 0. We demonstrate examples of noise introduction through background replacement and averaging.

At the end of training, our model was tested on a set of 100 previously unseen patients selected at random from the full dataset prior to training. As shown in Figure 5.5, when

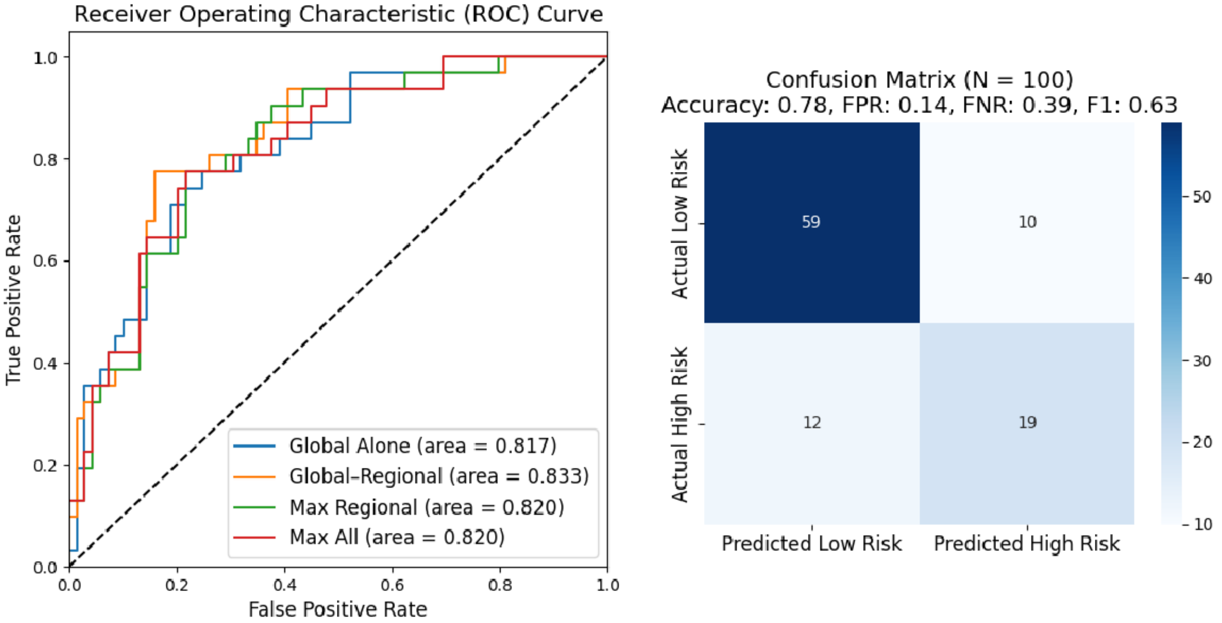

Figure 5.5: **Test Set Results.**(Left) Receiver operator characteristic curves are shown for four different sample evaluation methods. In the patient-level curve, we consider only the global prediction per patient. In the max. regional curve, we consider only the maximum among region-level predictions. In the patient-regional curve, we average the global prediction with the maximum region-level prediction. Finally, in the "max. all" curve, we consider the maximum among all predictions. (Right) Confusion matrix for predictions on our test set.

using the patient-level prediction alone, we achieved an AUROC of 0.817, on par with the AUROC reported by Howard et al. of 0.814. The average of the patient-level score and the maximum region-level score performed best, with an AUROC for this test set was 0.833. This metric produced an accuracy and F1 score of 0.78 and 0.633, respectively. In contrast to the original study, our model was able to generate risk assessments directly from patient slides without the need for tumor region annotations or clinical parameters such as patient age, tumor size, progesterone receptor (PR) status, tumor grade, and histologic subtype.

5.4.3 Visualizing High-Risk Histologic Patterns

A crucial step in our study was investigating which histological patterns were closely associated with higher risk scores. To this end, we designed a visualization tool to generate high-risk heatmaps over patient slides. The risk maps, shown in Figure 5.6, combined three separate maps to improve confidence in our chosen regions. The first map demonstrates regions in the slide that were more closely attended to during the processing steps in the vision transformer blocks. The second map, termed the drop-out map, was generated by randomly occluding certain regions in our slide. When high-risk regions in a patient slide are occluded, they produce lower risk scores from our model. The third map was a coarse

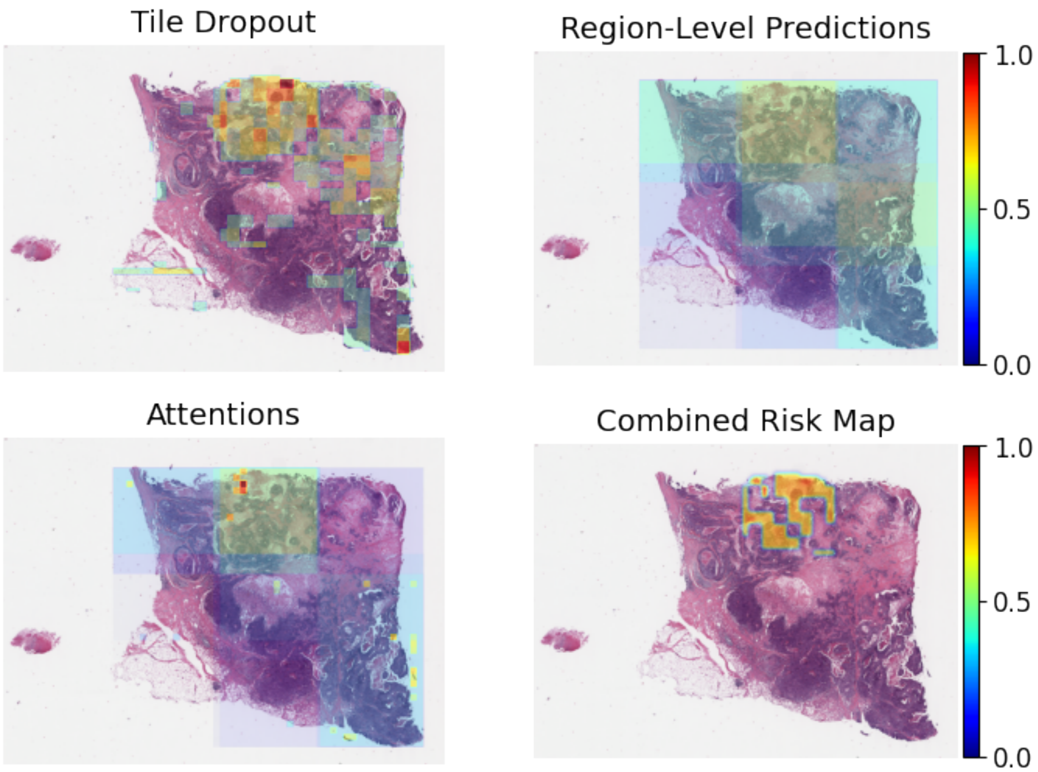

Figure 5.6: **Visualizing High-Risk Regions.**(Top Left) Tile Dropout Map. This heatmap demonstrates regions in the slide that, when occluded, resulted in lower model predictions. (Top Right) Region-level predictions. Region-level predictions produced by our model were mapped over the same slide. (Bottom Left) Attention Map. This map combines the attention weights produced by the transformer blocks with the region-level attention weights produced when synthesizing data for a single patient. (Bottom Right). Combined Risk Map. This heatmap combined the top percentiles of each map to create a final heat map of high-risk areas.

risk map generated from the region-level risk predictions $\widetilde{r}_{i,\ j}$. As can be seen in the example in Figure 5.6, all three maps generally coincided with each other. However, it was necessary to combine all three maps to remove spurious high-risk zones in the attention and dropout maps. In the attention maps, these zones may have corresponded to regions where the model paid close attention but deemed unimportant to the risk assessment. In dropout maps, these false alarm zones could have occurred when low-risk tiles were grouped together with high-risk tiles in a single occlusion trial.

We selected the 19 patients correctly predicted to have a high-risk score in our test set. A combined risk heatmap was generated for each patient, and the top 0.5% of tiles were selected from the maps. A total of 3,645 tiles were selected and encoded using our pre-trained image encoder. These high-risk representations were then clustered together and visualized in two dimensions using the tSNE dimensionality reduction algorithm, as shown in Figure 5.7A. Five of the 26 discovered clusters are shown in Figure 5.7B, demonstrating examples of nuclear pleomorphism, necrosis, and high cellularity.

We further evaluated our high-risk tiles by processing them with the tumor detection model introduced in Chapter 4 (Figure 5.7C). As previously noted, in their foundational work, Howard et al. generated patch-level risk predictions, which were then weighted by their likelihood of being in the tumor region. In our assessment, only 5 of 26 clusters produce an average tumor likelihood score below 75%, demonstrating that our model is focusing on the appropriate regions of the slide. As a control, we also show a random cluster taken from known benign tissue in the annotated BreAst Cancer Histology images (BACH) dataset [42]. This cluster was shown to have an average tumor likelihood of 0.38%. We demonstrate the concordance between expert-annotated tumor regions and model-predicted regions in Figure 5.8. In this example, we also note that only a small portion of the tumor region is considered high-risk, indicating that our model is picking up on patterns beyond those necessary to identify malignant regions.

5.5 Conclusions

In this chapter, we outlined a computational pipeline that predicts recurrence risk directly from digitized histology slides of early-stage breast cancer, thereby offering a cost-effective alternative to traditional genetic assays. Furthermore, the development of innovative visual tools led to the identification of high-risk slide regions, thereby enhancing the clinical utility and transparency of this AI-assisted prognostic tool.

Our algorithm identified high-risk regions that correlated with areas of high tumor likelihood, lending credibility to the model's decision-making ability. However, many areas of high tumor likelihood were not identified as high-risk regions, and some clusters of high-risk tiles had no apparent distinguishing features. In future work, we aim to increase the number of high-risk tiles discovered by expanding the test set, hence improving our capacity to build meaningful clusters. Beyond this, expert histopathologic annotations and advanced staining techniques like immunohistochemistry and FISH could be integrated with pre-trained tile embeddings to improve our clustering precision.

This ongoing study aims to establish more sophisticated and methodological approaches toward extracting the "morphologic biomarkers" of high recurrence risk. To evaluate our model's real-world performance, we plan to collect and analyze clinical records from Massachusetts General Hospital. Incorporating additional clinical metrics like those included in the Magee Equations™ could also improve the overall accuracy of the prognostic assessment and lower the possibility of overtreatment.

This technology has the potential to improve access to robust prognostic tools without the need for genetic testing and relieve the psychological and physiological burdens associated with the uncertainty surrounding cancer recurrence. As we continue this work, we hope to democratize high-quality care by improving clinicians' understanding of the morphological signatures associated with recurrence risk. This work exemplifies the dual role deep learning can play when it is integrated into our healthcare systems: both learning from and helping to expand clinical knowledge.

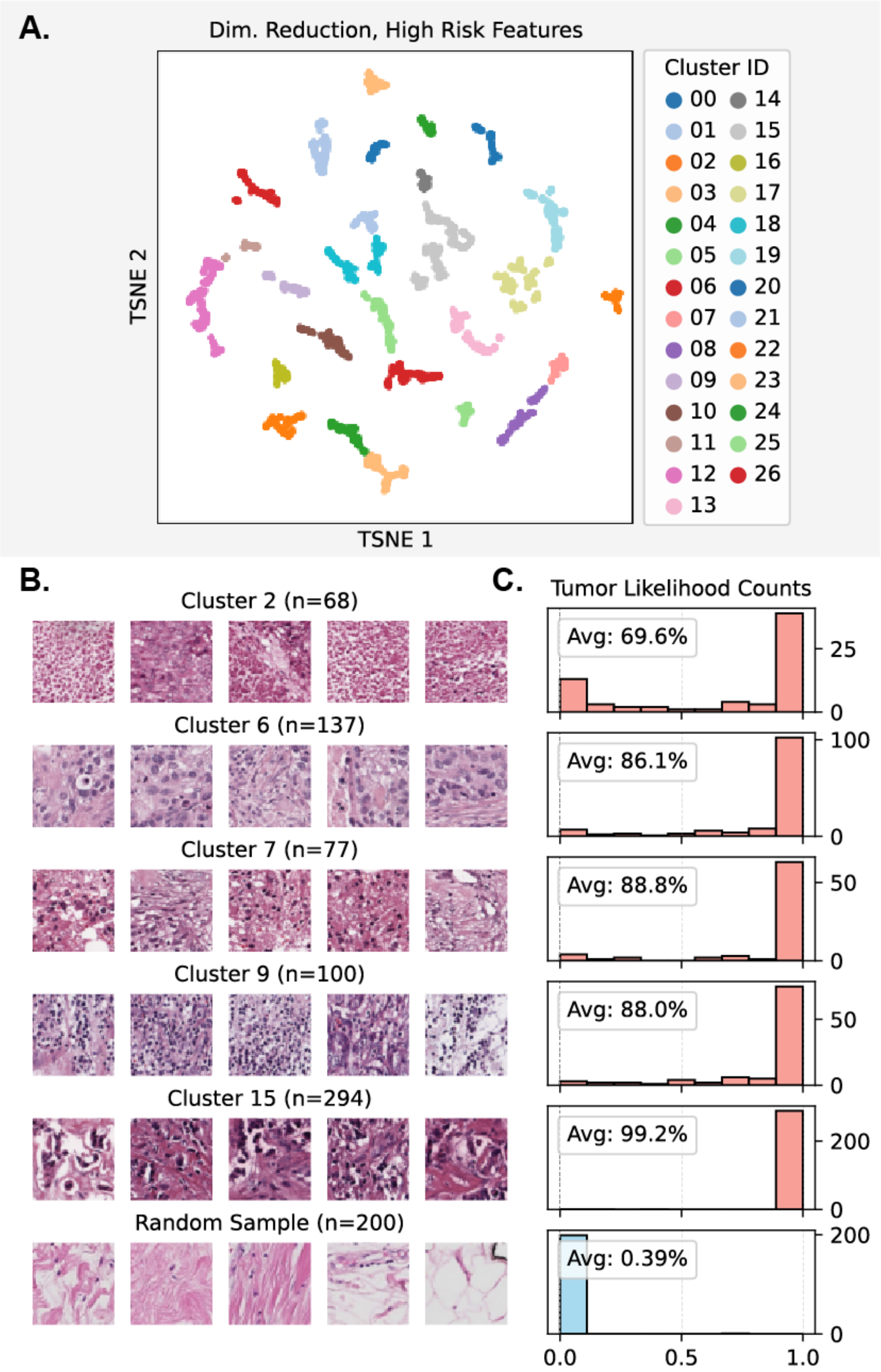

Figure 5.7: **Clustering High-Risk Tiles.**(A) tSNE dimensionality reduction for high-risk tile embeddings. Clusters are generated by the HDSCAN algorithm on full-length embeddings. (B) Clusters in our high-risk embeddings were converted to their original image formats for morphologic visualization. A random sample of benign tissue was also shown in the last row. (C) Corresponding histograms of the tumor likelihood of each tile in a specific cluster. The histogram for the random sample of normal tissue is shown in blue. The average tumor likelihood is noted for each cluster.

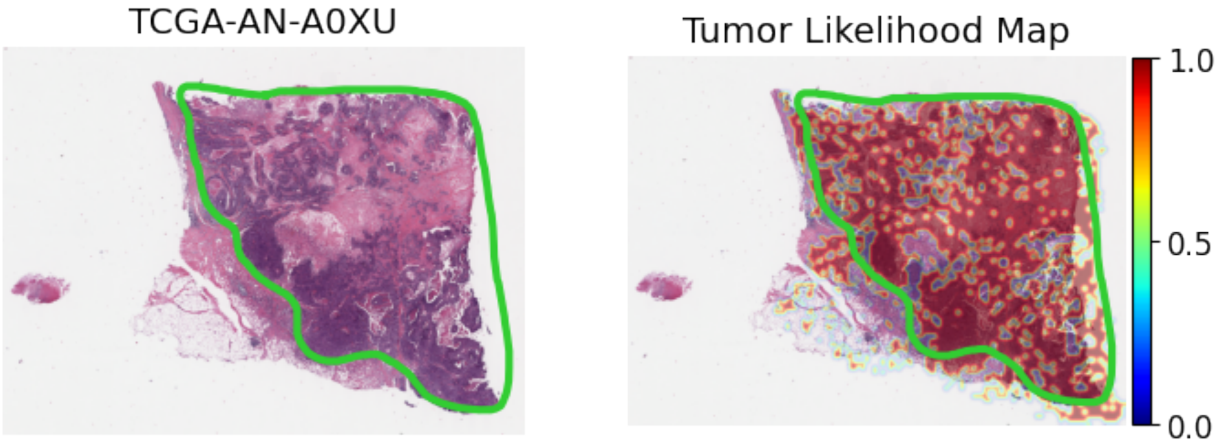

Figure 5.8: **Tumor Regions in Slide Example.**We demonstrate an example slide from TCGA (left), as well as an overlay of a tumor likelihood map (right). This map was constructed by evaluating each (256 × 256) tile in the slide with a pre-trained tumor-detection model. A rough expert annotation is shown in green[109].