Highlights

H&E to Oncotype DX: Predicting Recurrence Risk in HR+/HER2– Breast Cancer

Onur Can Koyun, Yongxin Guo, Ziyu Su, Hao Lu, Mostafa Rezapour, Robert Wesolowski, Gary Tozbikian, M. Khalid Khan Niazi, Metin N. Gurcan

- High Predictive Accuracy: Superior performance with AUCs of 0.89 and 0.86.
- Generalizability: Validated on a public multi-center dataset (TCGA) and an institutional dataset (OSU).
- Enhanced Interpretability: Identifies key biomarkers like LVI and Comedonecrosis.
- Cost and Time Efficiency: Reduces need for costly and time-consuming genomic assays.

H&E to Oncotype DX: Predicting Recurrence Risk in HR+/HER2– Breast Cancer

Onur Can Koyun[a], Yongxin Guo[a], Ziyu Su[b], Hao Lu[a], Mostafa Rezapour[c], Robert Wesolowski[d], Gary Tozbikian[e], M. Khalid Khan Niazi[b], Metin N. Gurcan[a]

[a] *Center for Artificial Intelligence Research, Wake Forest University School of Medicine, Winston-Salem, NC, USA*
[b] *Wexner Medical Center, The Ohio State University, Columbus, OH, USA*
[c] *Wake Forest Institute for Regenerative Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, USA*
[d] *Division of Medical Oncology, Comprehensive Cancer Center, The Ohio State University College of Medicine, Columbus, OH, USA*
[e] *Department of Pathology, The Ohio State University, Columbus, OH, USA*

Abstract

The Oncotype DX® Recurrence Score (ODX) is a clinically validated prognostic assay for hormone receptor–positive early-stage breast cancer. Despite its utility, widespread adoption remains limited due to high costs and lengthy turnaround times. In this study, we propose AnchorMIL, a novel framework that predicts ODX scores directly from H&E stained whole-slide images, with the potential to decrease dependence on molecular profiling. AnchorMIL employs an anchored regression-classification mechanism to predict both continuous risk scores and binary risk predictions. On the TCGA-BRCA and OSU cohort, AnchorMIL achieved AUCs of 0.89 and 0.86, respectively, Beyond predictive accuracy, AnchorMIL demonstrates promising generalizability, and its interpretability reveals biologically meaningful features of aggressive tumor biology. The model captures complex prognostic interactions, and prioritizes dominant risk factors in high-risk cases. AnchorMIL offers a scalable, cost-effective tool for risk stratification, with the potential to reduce reliance on genomic assays, accelerate treatment decisions, and support equitable breast cancer care.

1. Introduction

Breast cancer ranks as the second leading cause of cancer-related mortality among women in the United States, with approximately 316,000 new diagnoses and nearly 43,000 fatalities annually [1]. Survival benefits have been observed in patients with early-stage breast cancer who undergo systemic anti-cancer therapy following curative resection [2]. Clinically, treatment and prognosis are primarily determined by molecular subtype classification, which is based on the expression of estrogen, progesterone, and Human Epidermal Growth Factor Receptor 2 (HER2). This classification system delineates breast cancers into four distinct subtypes: hormone receptor–positive (HR+), HER2-negative, HER2-positive, and triple-negative. Notably, the HR+ and HER2-negative subtype constitutes approximately 70% of all cases [3, 4].

The Oncotype DX recurrence score (ODX), introduced by Genomic Health, has improved personalized therapeutic decision-making for breast cancer patients [5, 6]. This assay quantitatively evaluates the risk of cancer recurrence by analyzing the expression patterns of specific genes within tumor samples. Specifically, the test produces a recurrence score ranging from 0 to 100, with higher values indicating greater risk. Scores above 25 are considered high-risk, warranting the addition of chemotherapy to endocrine therapy, whereas scores below this threshold suggest a low recurrence risk that can typically be managed effectively with endocrine therapy alone [7, 6]. Extensive clinical trials confirm that patients classified with low ODX scores achieve excellent prognoses without the added burden of chemotherapy, underscoring the test's clinical value in preventing overtreatment [7, 8]. Consequently, contemporary clinical guidelines strongly endorse incorporating the ODX score into routine clinical practice to optimize therapeutic approaches [7]. Nevertheless, despite its proven clinical utility, the widespread adoption of the ODX test faces challenges related to cost-effectiveness and accessibility, given its high price and limited availability in certain healthcare settings [5].

Advances in digital pathology and artificial intelligence have significantly transformed diagnostic and prognostic practices, with whole slide image (WSI) analysis emerging as a key tool. This technology facilitates automated examination of complete pathological slides, improving diagnostic accuracy across various conditions [9, 10]. Nonetheless, two principal challenges persist in digital pathology: the gigapixel resolution of WSIs, which exceeds current hardware capabilities, and the scarcity of tissue-level annotations, which weakens the correlation between diagnostic labels and WSIs. These challenges

impede supervised learning and have led to the adoption of weakly supervised methods, among which Multiple Instance Learning (MIL) has shown promise for WSI classification [11, 12] For instance, recent work such as CASIIMIL has achieved an AUC of 0.96 on an NSCLC dataset for lung cancer subtyping through prototype learning [12].

When applying MIL approaches to predict ODX scores [13, 14, 15], current methodologies still exhibit performance limitations. Although Su et al. [12] reported an AUC of 0.86 for breast cancer pathology image classification across various ODX score ranges, their study was constrained by a small dataset of 150 cases, including only 18 test slides. Such limitations restrict comprehensive model validation and diminish the generalizability of the findings to broader clinical contexts. These challenges emphasize the need for larger, more representative datasets and enhanced MIL strategies to improve the accuracy and clinical applicability of ODX score prediction models.

Recent advances have seen the emergence of multimodal approaches for predicting Oncotype DX score. For example, Orpheus[16] leverages a vision–language framework to jointly analyze H&E-stained whole-slide images and their associated pathology reports, achieving an AUC of 0.89. However, its performance depends heavily on synthetic datasets to augment the limited real-world samples. Current approaches for predicting Oncotype DX (ODX) risk scores predominantly rely on classification methods despite their inherently continuous nature. In this paper, we reframe the problem as a joint regression and classification task and introduce AnchorMIL, a novel deep learning framework designed to predict ODX risk scores from WSIs. The core innovation of AnchorMIL is its anchored regression mechanism, which reimagines biomarker prediction. Instead of a monolithic regression model that learns a single global function, our framework delegates the prediction to multiple, specialized anchors. Each anchor learns to identify morphological features corresponding to a specific range of the biomarker score, effectively acting as a local expert for a distinct risk level. This distributed structure provides a more robust and granular prediction, as the final score is a confidence-weighted consensus derived from these expert opinions, ensuring the model's decision is supported by localized evidence.

In contrast to methods that frame the task as pure classification, our anchored regression scheme enables highly accurate continuous risk score predictions while also facilitating robust high-/low-risk classification through joint optimization with a binary cross-entropy loss. By unifying regression and classification within a single architecture, AnchorMIL captures the continuous

nature of recurrence scores while supporting clinically actionable decision thresholds. Furthermore, the anchored design enhances interpretability by associating specific morphological patterns with distinct risk levels, which the attention mechanism can highlight. Together, these innovations make AnchorMIL a comprehensive framework with the potential for clinical relevance for scalable risk stratification in early-stage breast cancer.

2. Related Work

The field of histopathological image analysis has witnessed significant progress with the application of MIL to WSIs, which are characterized by their gigapixel dimensions and the inherent challenge of limited detailed annotations. In this framework, each WSI is modeled as a "bag" containing numerous "instances" (patches), with only a bag-level label provided [17]. In other words, each gigapixel WSI-despite containing vast amounts of information-is annotated with a single label determined by the specific task at hand. This weakly supervised learning paradigm has been instrumental in harnessing the rich information embedded within WSIs without relying on exhaustive pixel-level or region-level annotations.

Initial approaches often relied on aggregating hand-crafted features through simple pooling techniques to form a global slide representation. While these methods laid the groundwork for MIL in medical imaging, their ability to capture tumor tissues' complex and heterogeneous nature was limited. The advent of attention-based MIL models provided a substantial improvement by explicitly weighting the contribution of individual instances. For example, the ABMIL model introduced an attention mechanism that assigns importance scores to patches, thereby focusing on regions most critical to the final prediction [18]. Similarly, the CLAM framework advanced the field by using instance-level clustering to both aggregate features and identify diagnostically relevant regions within WSIs, with several subsequent variants refining its methodology [19].

Recent work has pushed the frontier by incorporating more expressive deep learning models for feature aggregation. Transformer-based architectures, such as those employed in TransMIL, have been utilized to capture correlated information across instances. Notably, TransMIL has reported an impressive AUC of up to 98.82% in binary tumor classification on the TCGA-RCC dataset, demonstrating the potential of leveraging self-attention mechanisms to model long-range dependencies among patches [20, 21]. Furthermore,

models like Mamba [22] have also been explored within MIL frameworks, emphasizing scalability and enhanced feature representation.

Beyond the enhancements achieved by transformer architectures, recent advancements have turned toward prototype learning to improve MIL performance. Early works in this area, including CASIIMIL and ProtoMIL, introduced the concept of integrating prototypes to enhance sensitivity to small tumors and to guide the learning process [23, 24]. CASIIMIL, for instance, employs negative representation learning to adjust the model's sensitivity toward minor but clinically significant features. In contrast, ProtoMIL constructs prototypes through similarity calculations and leverages attention scores to focus on diagnostically important regions. Despite their merits, these approaches have so far incorporated prototypes primarily as preliminary guidance rather than as a core component throughout the model architecture and learning process.

Additional efforts in MIL have sought to address the challenges posed by the enormous number of instances per slide. A double-tier MIL framework utilizing pseudo bags was developed to manage the limited number of bags relative to the plethora of patches, qualitatively demonstrating that derived instance probabilities could outperform standard attention scores for positive region detection [25]. Similarly, DeepSMILE integrates a feature variability-aware variant of DeepMIL with a self-supervised, histopathology-specific feature extractor (VarMIL) to effectively model intra-tumor heterogeneity, achieving state-of-the-art performance on datasets such as CAMELYON-16 and improving classification on lung and other cancers [26, 27].

Another notable advancement is the DAS-MIL framework, which brings a multi-scale perspective by enabling information flow across different resolutions in a pyramidal WSI structure using message passing. Additionally, DAS-MIL incorporates a knowledge distillation scheme to align latent representations across resolutions while retaining informative diversity, thus enhancing the robustness of slide-level predictions [28].

The utilization of Vision Transformers has further enriched MIL methodologies in medical imaging. Recent studies have explored the application of transformer architectures in this domain, discussing their core components, attention mechanisms, and inherent limitations. These works also underscore the challenges posed by diverse learning paradigms in medical imaging, setting the stage for future research that could unify multi-scale and prototype-based methods under a more cohesive framework [29].

WSI analysis have evolved from pooling and hand-crafted feature aggre-

Table 1: Summary of Multiple Instance Learning methods for WSI analysis

Method	Datasets	Primary Task	# Slides	AUC	Limitations
Orpheus [16]	MSK/IEO/MDX-BRCA (6172 cases)	RS prediction + recurrence	6172	RS > 25: 0.89; RS ≤ 25: 0.75 (time-AUC)	Requires multimodal input (H&E + report); performance drops on RS 11–25; complex setup
TransMIL [21]	CAMELYON16; TCGA-NSCLC; TCGA-RCC	Binary/multi-class WSI classification	400; 993; 884	0.93; 0.96; 0.99	High memory usage; needs Nyström approximation; evaluated only at 20×
ProtoMIL [24]	Bisque; Camelyon16; TCGA; Colon; Messidor	MIL for WSI classification	58–1016	0.89–0.96	Prototype meaning unclear; vulnerable to adversarial attacks; low performance on small datasets
DTFD-MIL [25]	CAMELYON16; TCGA-Lung (patch-level)	WSI-based cancer diagnosis	941	0.95 (AFS); 0.96 (MaxMinS)	Pseudo-bag noise; excessive pseudo-bags hurt performance; distillation strategy sensitive
DeepSMILE [26]	TCGA-CRC; TCGA-BRCA	MSI (CRC); HRD (BRCA) classification	360; 1127	0.86 (MSI); 0.81 (HRD)	Label quality sensitive (e.g., HRD thresholds); no external validation; lacks spatial modeling; depends on SSL quality
Deep-ODX [12]	151 HR+/HER2- breast cancer WSIs (tumor regions)	ODX recurrence risk prediction (high vs. low)	151	0.862 ± 0.034	Moderate sensitivity near ODX = 25; higher errors for scores 20–30; manual tumor annotation required
DAS-MIL [28]	CAMELYON16; TCGA-Lung (LUAD + LUSC)	Tumor subtype classification with MIL	400; 1054	0.973 (Camelyon); 0.965 (Lung)	Complex GNN-distillation setup; relies on pre-extracted features (e.g., DINO); not end-to-end
CASiiMIL [23]	CAMELYON16; CAMELYON17	Tumor vs. normal WSI classification	399; 1000	0.9679 (C16); 0.9446 (C17, Center 3)	High compute due to all negative keys; requires efficient key set pruning to reduce redundancy
CLAM [19]	TCGA; CAMELYON16/17	Subtyping; LN metastasis detection	3750	RCC: 0.991; NSCLC: 0.956; LN: 0.953	Not optimized for patch-level localization; requires careful instance-level label tuning

gation methods to sophisticated models that integrate attention mechanisms, transformer architectures, and prototype learning. Our work introduces a novel framework anchored on the joint regression and prediction of the ODX score; a vital biomarker for assessing tumor aggressiveness and patient prognosis. In contrast to earlier methods that largely approached histopathological analysis as a classification problem, our approach integrates continuous risk estimation through regression alongside traditional diagnostic predictions. By leveraging state-of-the-art MIL techniques including cross attention-based models and transformer architectures, we aim to capture both the intricate spatial heterogeneity and subtle morphological nuances within WSIs. This integrative strategy bridges the gap between molecular assays and image-based diagnostics and provides a more precise, clinically interpretable quantification of tumor behavior, ultimately enhancing the potential for tailored therapeutic interventions. Table 1 presents an overview of MIL methods, highlighting their primary tasks, AUC scores, and associated limitations.

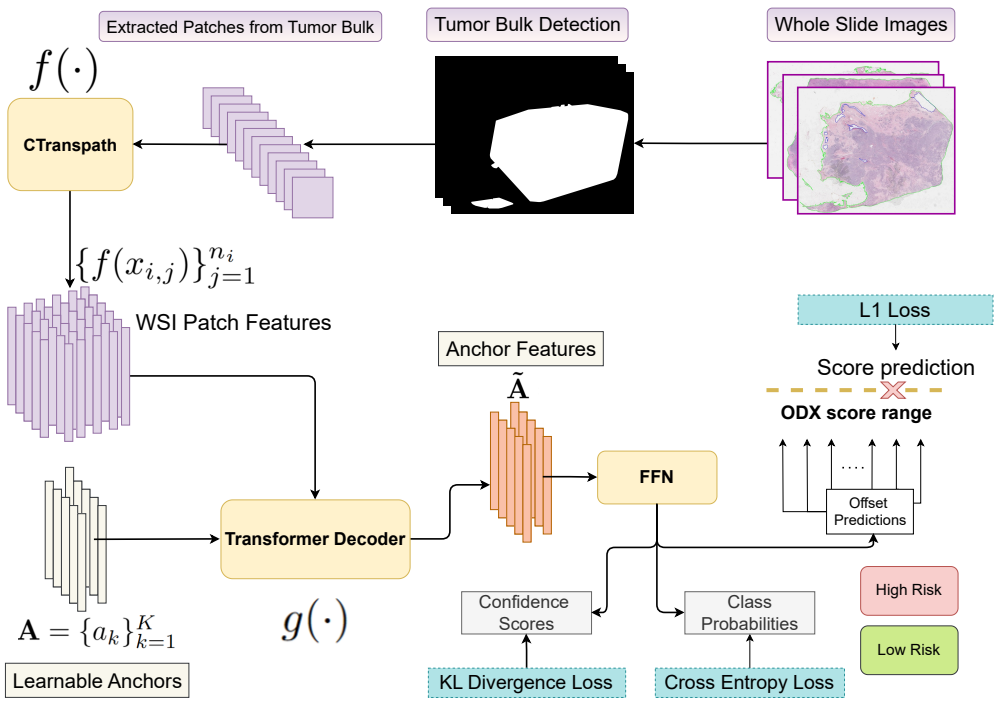


Figure 1: Overview of AnchorMIL framework. After detecting the tumor bulk, H&E regions are split into patches and embedded via CTransPath. Learnable anchors decode these features through cross-attention, producing representations that drive parallel heads for continuous recurrence-score regression, high/low-risk classification, and per-anchor confidence.

3. Methodology

In this work, we introduce AnchorMIL, a transformer-based deep learning framework for whole-slide image (WSI)-driven prediction of Oncotype DX (ODX) recurrence scores, formulated as both a continuous regression and a binary risk classification task. As illustrated in Fig. 1, AnchorMIL operates through four sequential stages.

First, shown in Fig 3 tumor bulk detection is performed to isolate the most informative regions of each WSI. This is achieved using a fine-tuned ResNet50-based CNN scorer, which evaluates patches extracted from annotated tumor regions and assigns each a discriminative score reflecting its relevance to recurrence risk. Rather than applying a fixed threshold, the top-ranked patches are selected based on their scores, ensuring that only the most representative tumor regions are retained for further analysis.

Second, the selected patches are processed by a shared encoder to generate instance-level feature embeddings. These embeddings are then passed to a transformer decoder, where a set of learnable anchor vectors interact with the patch features. Through cross-attention, the model aggregates contextual information from the tissue, while self-attention captures relationships among the anchors themselves, resulting in refined anchor representations.

In the third stage, each refined anchor is independently mapped to three outputs: a confidence weight indicating its reliability, a regression offset that adjusts a predefined anchor score, and a classification probability estimating the likelihood of high-risk recurrence. The architecture of the proposed method is illustrated in Fig. 2.

Finally, the outputs from each anchor are combined using their respective confidence weights to generate a continuous prediction of the ODX recurrence score, along with an aggregated probability for binary risk classification. By anchoring predictions to interpretable score intervals and modulating them through learned confidences, AnchorMIL delivers both high predictive accuracy and enhanced interpretability in breast cancer risk stratification.

3.1. Multiple Instance Learning Paradigm

In this work, we adopt an MIL framework for risk stratification from WSIs of breast cancer and extend it with an anchor-based regression and classification model. Let the dataset be denoted as

$$\mathcal{X} = \{X_1, X_2, \ldots, X_N\},$$

where each X_i represents a WSI. Each WSI is partitioned into a set of non-overlapping patches:

$$X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}\},$$

with n_i being the number of patches extracted from the i-th slide. In the context of MIL, each WSI is treated as a *bag* and the individual patches as *instances* within that bag.

Traditionally, MIL assumes that a positive bag contains at least one positive instance while a negative bag contains only negative instances. However, for our problem, the overall ODX score of a bag is derived from the collective contributions of all instances rather than from isolated high-risk patches. Each WSI is associated with a binary risk label $Y_i \in \{0, 1\}$ (low-risk: 0, high-risk: 1), based on clinical criteria.

Let $f(\cdot)$ denote a patch-level feature encoder, and $g(\cdot)$ denote an aggregation function that synthesizes patch-level features to yield a slide-level representation. The relationship between instance-level features and the final risk label is formalized as:

$$Y = g\Big(\{f(x_{i,j})\}_{j=1}^{n_i}\Big), \quad i = 1, \ldots, N.$$

3.2. Anchor-based Regression and Classification for ODX Score Prediction

To enhance the standard MIL framework, we propose an anchor-based model that jointly performs regression and classification to predict the continuous ODX score. In our model, the *anchors* are implemented as learnable vectors that are optimized during training to predict ODX scores. These anchors are designed to effectively aggregate information from the patch-level features, enabling them to focus on the most informative regions of the WSI.

Let $\mathbf{A} = \{a_k\}_{k=1}^{K}$ denote the set of learnable anchor vectors, where K is the total number of anchors. Each anchor a_k interacts with the patch feature representations $\{f(x_{i,j})\}_{j=1}^{n_i}$ extracted by the encoder $f(\cdot)$, thereby gathering contextual information from the patches. This process produces an enhanced feature representation for each anchor, denoted as:

$$\tilde{\mathbf{A}} = \{\tilde{a_k}\}_{k=1}^{K} = g\Big(\{f(x_{i,j})\}_{j=1}^{n_i}, \{a_k\}_{k=1}^{K})\Big)$$

Where $g(\cdot)$ is a transformer decoder architecture that consist of cross-attention and self-attention layers. Subsequently, each anchor produces three key outputs:

1. Anchor Confidence: A score c_k indicating the reliability or importance of the anchor in contributing to the final risk prediction.
2. Regression Output: A continuous value r_k representing the predicted ODX score.
3. Classification Output: A probability p_k (or binary outcome) indicating the predicted risk status (low-risk or high-risk).

These outputs are computed using linear layers applied to the output anchor features:

$$c_k = h_c(\tilde{a_k}),$$
$$r_k = s_k + h_r(\tilde{a_k}),$$
$$p_k = h_p(\tilde{a_k}).$$

where h denotes the linear projection for each output, s_k represents the score associated with the k^{th} fixed anchor, and $h_r(\tilde{a_k})$ is the predicted regression offset . The regression output r_k provides a continuous estimate of the ODX score for each anchor, capturing fine-grained risk variations. In contrast, the classification output p_k delivers a binary risk prediction that categorizes the WSI into low-risk or high-risk based on the threshold defined in TAILORx. The anchored regression approach allows the model to benefit from both continuous and categorical perspectives, thereby enhancing the overall risk prediction accuracy.

This integrated method not only leverages the discriminative power of the individual patches but also effectively aggregates their contributions through the anchors. As a result, the model delivers robust predictions by combining the detailed information captured in the regression outputs with the decisiveness of the classification outputs.

3.3. Anchored Regression with Fixed Anchor Scores and Maximum Confidence Selection

After extracting a rich feature representation through the Transformer module, an anchored regression head predicts the continuous ODX risk score. We introduce a set of fixed anchor scores, $\{s_k\}_{k=1}^{K}$, which are spaced evenly over the ODX score interval. For each fixed value s_k, the model predicts:

$$r_k = s_k + h_r(\tilde{a_k}),$$

where $h_r(\tilde{a_k})$ is the offset learned by the network. Each anchor is also associated with a confidence score c_i that reflects its reliability in predicting the risk score.

For the final prediction, instead of selecting the anchor with the highest confidence score, we aggregate the outputs of all anchors by taking a weighted sum where each regression output is multiplied by its associated confidence. Specifically, the final continuous ODX risk score prediction is defined as

$$\hat{r} = \sum_{i=k}^{K} c_k \, r_k,$$

where c_k is the confidence score and r_k is the regression output for anchor k.

Similarly, if a binary risk classification is required, we compute an aggregated classification output as

$$\hat{p} = \sum_{k=1}^{K} c_k \, p_k,$$

and determine the final risk label by comparing the aggregated probability to a predefined threshold τ:

$$Y = \begin{cases} 1, & \text{if } \hat{p} > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

This approach, which combines anchors with learned offsets and confidence-based aggregation, provides robust and interpretable predictions of the ODX risk score by accounting for each anchor's contribution proportionally to its confidence.

3.4. Network Architecture and Attention Mechanisms

The proposed network architecture builds upon a Transformer-based backbone, incorporating three core blocks in a residual manner: (i) *Cross-Attention*, (ii) *Self-Attention*, and (iii) a *Feed Forward* block. This design allows efficient aggregation of local and global contextual cues across multiple patches extracted from the WSI.

Cross-Attention.. Let $\mathbf{X} \in \mathbb{R}^{B \times N \times d}$ be the patch features, where B is the batch size, N is the number of patches, and d is the feature dimension, and let $\mathbf{A} \in \mathbb{R}^{K \times d}$ be the learnable anchors, where K is the number of anchors. We first compute the query from the anchors and the key–value pairs from the patch features:

$$\mathbf{Q} = \mathrm{LN}(\mathbf{A})\, \mathbf{W}_Q, \quad [\mathbf{K},\, \mathbf{V}] = \mathrm{LN}(\mathbf{X})\, \mathbf{W}_{KV},$$

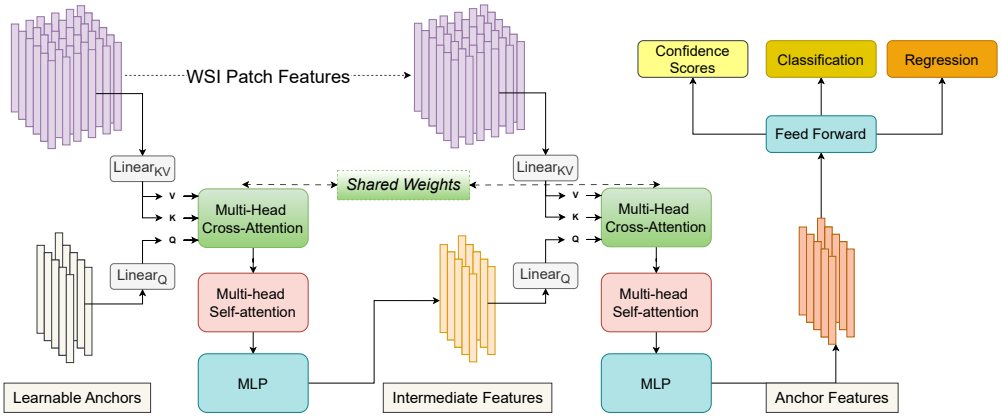


Figure 2: Overview of the Transformer decoder architecture. In this design, cross-attention weights are shared across layers, while all other layers maintain independent parameters. The output anchor features are subsequently fed into three distinct heads: a classifier, a regression module, and a confidence predictor.

where $\mathbf{W}_Q \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_{KV} \in \mathbb{R}^{d\times 2d}$. Here, $\mathrm{LN}(\cdot)$ denotes layer normalization and $\mathrm{split}(\cdot)$ partitions the output along the feature dimension into two tensors $\mathbf{K}$ and $\mathbf{V}$ each in $\mathbb{R}^{B\times N\times d}$.

For multi-head attention, we divide the feature dimension into H heads with $d_h = d/H$. Let $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$ denote the projection of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ onto head i, respectively. For each head, the scaled dot-product attention is computed as

$$\mathrm{head}_i = \mathrm{softmax}\left(\frac{\mathbf{Q}_i\,\mathbf{K}_i^\top}{\sqrt{d_h}}\right)\mathbf{V}_i.$$

The multi-head attention output is obtained by concatenating the outputs of all heads and applying a final linear projection:

$$\mathrm{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Concat}\Big(\mathrm{head}_1,\ldots,\mathrm{head}_H\Big)\,\mathbf{W}_O,$$

with $\mathbf{W}_O \in \mathbb{R}^{d\times d}$. Finally, we add a residual connection to preserve the original query:

$$\mathbf{A}^* = \mathbf{Q} + \mathrm{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}).$$

Self-Attention.. Let $\mathbf{A}^*$ be the anchor features output of the cross-attention layer. We compute the query, key, and value from $\mathbf{A}^*$ as

$$[\mathbf{Q_{SA}},\,\mathbf{K_{SA}},\,\mathbf{V_{SA}}] = \mathrm{LN}(\mathbf{A}^*)\,\mathbf{W}_{QKV_{SA}},$$

where $\mathbf{W}_{QKV_{SA}} \in \mathbb{R}^{d\times 3d}$ and output partitions into three tensors of shape $\mathbb{R}^{B\times N\times d}$.

For multi-head self-attention, we divide the feature dimension into H heads with $d_h = d/H$. Let $\mathbf{Q_{SA}}_i$, $\mathbf{K_{SA}}_i$, and $\mathbf{V_{SA}}_i$ denote the projections onto head i. For each head, the scaled dot-product attention is computed as

$$\text{head}_{\text{SA}i} = \text{softmax}\left(\frac{\mathbf{Q_{SA}}_i \mathbf{K_{SA}}_i^\top}{\sqrt{d_h}}\right)\mathbf{V_{SA}}_i.$$

The multi-head self-attention output is then obtained by concatenating the head outputs and applying a final linear projection:

$$\text{MHA}(\mathbf{Q_{SA}}, \mathbf{K_{SA}}, \mathbf{V_{SA}}) = \text{Concat}\Big(\text{head}_{\text{SA}1}, \ldots, \text{head}_{\text{SA}H}\Big)\,\mathbf{W_{O_{SA}}},$$

with $\mathbf{W}_{O_{SA}} \in \mathbb{R}^{d\times d}$. Finally, adding a residual connection yields

$$\tilde{\mathbf{A}} = \mathbf{A}^* + \text{MHA}(\mathbf{Q_{SA}}, \mathbf{K_{SA}}, \mathbf{V_{SA}}).$$

Feed Forward.. Following cross attention and self-attention blocks, a two feed-forward layer (FF) with GELU activation in between is applied:

$$\tilde{\mathbf{A}} = \text{FF}\big(\text{GELU}(\text{FF}(\tilde{\mathbf{A}}))\big) + \tilde{\mathbf{A}}.$$

We construct our model by stacking two blocks consisting of cross-attention, self-attention, and MLP layers. In the cross-attention blocks, the weights are shared in the same manner as in the Perceiver architecture [30].

3.5. Loss Function

We adopted an anchored regression framework augmented by a binary cross-entropy (BCE) term to train the `AnchorMIL` model for both continuous ODX score prediction and high/low risk classification. Let

$$\mathbf{r} \in \mathbb{R}^{B\times K} \quad , \quad \mathbf{p} \in \mathbb{R}^{B\times K} \quad \text{and} \quad \mathbf{c} \in \mathbb{R}^{B\times K}$$

denote the per-anchor regression scores, binary predictions and the corresponding confidence scores, respectively, for a batch of size B and K anchors. We denoted the ground-truth scores by

$$\mathbf{y} \in \mathbb{R}^{B}.$$

1) Regression Loss.. We used a per-anchor L1 (mean absolute error) loss to measure how well each anchor-based prediction $r_{b,k}$ matches the ground-truth y_b where b is the mini-batch size and y_b is the continuous ODX scores in a mini-batch.

$$\text{errors}_{b,i} \;=\; \text{L1}\big(r_{b,k},\; y_b\big),$$

which is then averaged across both anchors and the batch:

$$\mathcal{L}_{\text{reg}} \;=\; \frac{1}{K}\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{K}|r_{b,i} - y_b|$$

By default, we may either take the mean over all anchor errors (unweighted) or weight each anchor's error by its predicted confidence $c_{b,k}$ if the application demands.

2) Confidence Matching Loss.. To guide the model in assigning higher confidence to anchors that better approximate the target, we first constructed a *target confidence* distribution $\tilde{c}_{b,k}$ for each sample b and anchor a_k. A simple strategy is to use a softmax function over the anchor-specific errors:

$$\tilde{c}_{b,k} = \frac{\exp\Big(-\log\Big((a_k - y_b)^2\Big)\Big)}{\sum_{k=1}^{K}\exp\Big(-\log\Big((a_k - y_b)^2\Big)\Big)},$$

where a_k is the fixed anchor value. In practice, using a `softmax` with logarithmic function encourages low-error anchors to receive higher target confidence. Softmax function yields a probability distribution over fixed anchors: $\tilde{c}_b$. We then apply the Kullback–Leibler (KL) divergence to match the predicted confidence distribution c_b to target confidence $\tilde{c}_b$:

$$\mathcal{L}_{\text{conf}} \;=\; \frac{1}{B}\sum_{b=1}^{B}\text{D}_{\text{KL}}\Big(\mathbf{c}_b \;\big\|\; \tilde{\mathbf{c}}_b\Big),$$

where each $\mathbf{c}_b$ satisfies $\sum_{k=1}^{K} c_{b,k} = 1$. This term encourages consistency between the model's confidence assignment and the anchor-specific regression errors.

3) Binary Classification Loss.. A binary label can be derived from $\mathbf{y}$ by thresholding at some value τ, *i.e.*,

$$y_{\text{bin}} \;=\; \begin{cases} 1, & \text{if } \; y > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

To obtain a final scalar prediction for classification, we formed a confidence-weighted prediction,

$$\hat{p}_b \;=\; \sum_{k}^{K} c_{b,k}\, p_{b,k},$$

and transformed it into a probability via a logistic function:

$$\hat{\pi}_b \;=\; \sigma\big(\hat{p}_b\big),$$

where $\sigma(\cdot)$ is the sigmoid function. The BCE loss is then

$$\mathcal{L}_{\text{BCE}} \;=\; \frac{1}{B}\sum_{b=1}^{B} \text{BCE}\big(\hat{\pi}_b,\; y_{\text{bin},b}\big).$$

This term aligns the continuous regression outputs with high/low risk labels, improving performance for classification tasks.

Total Loss.. The overall loss is a linear combination of the three terms:

$$\mathcal{L} \;=\; \beta\mathcal{L}_{\text{reg}} \;+\; \mathcal{L}_{\text{conf}} \;+\; \mathcal{L}_{\text{BCE}},$$

where β is a weighting hyperparameter controlling the emphasis on regression relative to binary classification. In practice, adjusting β offers a convenient way to shift priority between continuous score accuracy and discrete risk classification.

3.6. Tumor Bulk Detection

To identify tumor regions within whole slide images (WSIs), we employed a patch-wise scoring strategy using a pretrained and fine-tuned ResNet50-based [31] convolutional neural network (CNN), as described in BCR-Net [32]. This model assigns a discriminative score to each patch extracted from annotated tumor regions, enabling the selection of patches most indicative of tumor presence.

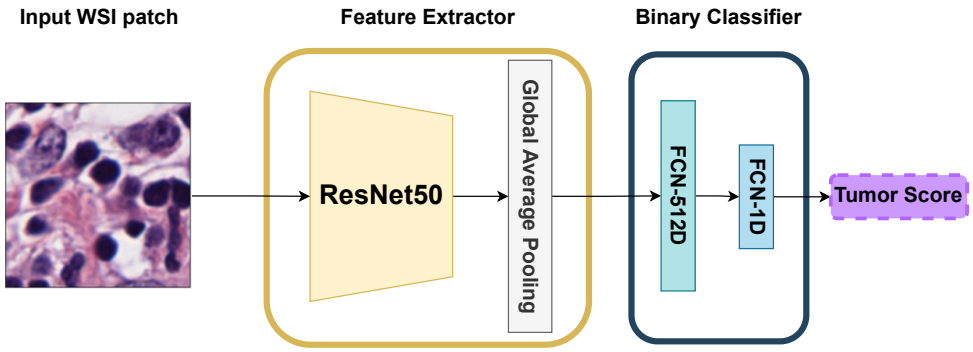


Figure 3: Architecture for patch-level tumor scoring. The model consists of a ResNet50 pretrained on ImageNet, followed by a global average pooling layer, two fully connected layers, and a sigmoid activation function. The output represents the probability of a patch being associated with tumor region, which is further used to compute the Tumor Score (TS) for intelligent patch selection.

To quantify the discriminative capacity of each patch, we computed a Tumor Score (TS) as follows:

$$\mathrm{TS} = \left|\sigma(f_\theta(x)) - 0.5\right| / 0.5, \tag{1}$$

where $f_\theta(x)$ denotes the CNN output for input patch x, and σ is the sigmoid activation function. The TS ranges from 0 to 1, with higher values indicating greater confidence in the patch's classification.

Rather than applying a fixed threshold, all patches within the annotated tumor region were scored and ranked in descending order based on their TS values. The top K patches with the highest scores were selected as representative of the tumor bulk. This intelligent sampling strategy ensures that only the most informative regions are retained for downstream analysis, reducing noise from non-discriminative or background tissue.

The selected patches were subsequently embedded into feature vectors using the same CNN-scorer (excluding the classification head), forming a bag of instances for slide-level classification via a multiple instance learning (MIL) model. This approach not only enhances classification performance but also significantly reduces computational cost, making it suitable for deployment in resource-constrained environments.

4. Experiments

4.1. Experimental Setup

The whole tissue regions in the original WSIs were segmented from the white background using Otsu's binarization algorithm. To mitigate staining variability, z-score normalization was applied to all extracted patches. For tumor bulk detection, each WSI was partitioned into non-overlapping patches of size 224×224 pixels at $20\times$ magnification. These patches were analyzed by the CNN-scorer, which employed a truncated ResNet50 backbone followed by a global pooling layer, two fully connected layers, and a sigmoid activation function. The sigmoid output represented the probability of each patch being classified as high-risk.

For feature extraction, tumor regions identified during the bulk detection stage were subsequently cropped at $20\times$ magnification into patches of size 896×896 pixels for both the TCGA-BRCA and OSU datasets.

The feature encoder $f(\cdot)$ was implemented using CTransPath [33], which was pre-trained on various histopathology datasets, yielding an output feature dimension of 768. Both the self-attention and cross-attention layers were configured with 12 heads. The model architecture consisted of 2 blocks; each block was composed of a multi-head cross-attention layer, a multi-head self-attention layer, and a multi-layer perceptron with GELU activation function in between. Notably, the weights for the multi-head cross-attention layers were shared across the two blocks. We selected 128 anchors as the final configuration. Various anchor counts ranging from 16 to 256 were evaluated by training on the TCGA-BRCA dataset, and 128 yielded the best performance.

During training, a mini-batch size of 32 was used. To handle the variable number of patches per WSI, the patches were sorted by their count, and the feature sequences were padded to match the maximum sequence length in each mini-batch. The model was optimized using the AdamW optimizer with an initial learning rate of 5×10^{-4} and a weight decay of 1×10^{-3}. A cosine learning rate scheduler was employed with a minimum learning rate of 1×10^{-5}. Training was performed for 20 epochs using PyTorch on a single NVIDIA A100 GPU.

4.2. Evaluation

At inference time, the aggregated probability prediction $\hat{p}$ and regression prediction $\hat{r}$ are used to classify the sample as high- or low-risk and predict continuous ODX score. We evaluate performance using metrics appropriate

Table 2: Summary of hyper-parameters and configuration details for AnchorMIL.

Parameter	**AnchorMIL Configuration**
Tumor Bulk Patch Size	224×224 pixels
Feature Patch Size	896×896 pixels
Magnification Level	$20\times$
Feature Extractor	CTransPath (pre-trained)
Feature Dimension	768
Number of Anchors	128
Number of Attention Heads	12
Number of Blocks	2
Shared Cross-Attention Weights	Yes
Activation Function	GELU
Dropout	0.1
Batch Size	32
Epochs	20
Optimizer	AdamW
Initial Learning Rate	0.0005
Learning Rate Scheduler	Cosine decay
Minimum Learning Rate	0.00001
Weight Decay	0.001
Warmup Epochs	5
β	0.7
GPU	NVIDIA A100

for both continuous and discrete predictions, including ROC-AUC, accuracy, F1 score, sensitivity, specificity, mean absolute error (MAE) and mean squared error (MSE).

4.3. Datasets

To rigorously assess the performance and generalizability of our proposed **AnchorMIL** framework, experimental validation was conducted on multiple breast cancer pathology datasets. These include an institutional cohort from The Ohio State University (OSU) and a publicly available dataset the TCGA-BRCA dataset each contributing distinct characteristics in terms of clinical relevance and image heterogeneity.

4.3.1. OSU Dataset

The OSU dataset comprises 398 hematoxylin and eosin (H&E) stained WSIs derived from $HR^+/HER2^-$ breast cancer specimens, collected at The Ohio State University Wexner Medical Center. Under an approved study protocol by the Ohio State University Cancer Institutional Review Board with all relevant consent and HIPAA waivers the slides were annotated with corresponding Oncotype DX (ODX) risk scores. Specifically, 267 WSIs were classified as low-risk (ODX$\leq$25) and 131 as high-risk (ODX$>$25). To address the inherent class imbalance, we employed a stratified sampling strategy. First, a balanced test set comprising 70 slides, with equal representation from both risk categories, was established. Subsequently, a three-fold cross-validation was conducted on the remaining cases, ensuring that all WSIs from the same patient were assigned to the same split to prevent data leakage.

Table 3: Demographics of HR+/HER2- patients in the TCGA-BRCA dataset. In this table, n indicates the total number, and SD indicates standard deviation.

		Risk Group	
		Low-risk	**High-risk**
Total, n		443	73
Age, mean (SD)		64.9 (9.8)	64.8 (9.4)
Sex, n (%)	Female	440 (99.3)	73 (100)
	Male	3 (0.7)	0
Race, n (%)	White	346 (78.1)	61 (77.2)
	Black	37 (8.4)	8 (10.1)
	Other	48 (10.8)	9 (11.4)
	Missing	12 (2.7)	1 (1.3)
Histologic type, n (%)	Lobular	86 (19.4)	23 (29.1)
	Ductal	294 (66.4)	57 (72.2)
	Lobular + Ductal	64 (14.2)	2 (2.5)
Grade, n (%)	1	44 (10.0)	3 (3.8)
	2	211 (47.6)	24 (30.4)
	3	188 (42.4)	48 (60.2)
	Missing	0	4 (5.1)

4.3.2. TCGA-BRCA Dataset

The TCGA-BRCA dataset is a widely recognized public resource for breast cancer research, originally containing 1,133 WSIs. Following stringent quality

control measures that excluded cases with incomplete receptor status, missing ODX scores, or processing failures, the final analytical cohort consisted of 1,065 WSIs from 1,006 unique patients. Within the $HR^+/HER2^-$ subgroup (n=516), risk stratification based on research-derived ODX scores computed using normalized mRNA expression data as described by Howard et al.[34] yielded 443 low-risk patients and 73 high-risk patients. To prevent data leakage., patient-level stratification was rigorously maintained to prevent data leakage; specifically, all WSIs belonging to the same patient were assigned to the same data split (training, validation, or test). Additionally, non-HR+/HER2– cases were incorporated into the training set to counteract class imbalance. High-risk cases are oversampled to boost their representation and balance the class distribution during training. After comprehensive data cleaning and three-fold partitioning, the dataset was divided into 563 WSIs for training, 80 for validation, and 213 for testing.

In summary, the OSU dataset contributes clinical authenticity and heterogeneity reflective of routine clinical practice, whereas the TCGA-BRCA dataset with its extensive clinical annotations serves as a publicly available benchmark.

4.4. Results

We conducted extensive experiments on both the TCGA-BRCA and OSU in-house datasets to evaluate the performance of our proposed **AnchorMIL** framework compared to existing state-of-the-art MIL-based methods. The primary performance metrics employed included Area Under the Receiver Operating Characteristic Curve (AUC), accuracy, F1-score, sensitivity, and specificity. Table 4 summarizes the comparative results on the OSU dataset. AnchorMIL demonstrated significant improvements over other methods, achieving the highest performance in AUC (0.858, 95% CI 0.87–0.90; $p<0.001$, DeLong's test vs. AUC=0.5), accuracy (0.800 ± 0.012), F1-score (0.790 ± 0.008), and specificity (0.848 ± 0.054). Notably, while ABMIL and S4 showed higher sensitivity (0.848 ± 0.082), AnchorMIL provided a superior balance between sensitivity and specificity, reflecting its robustness in accurately identifying both high-risk and low-risk cases.

The evaluation on the TCGA-BRCA dataset further validated the efficacy of AnchorMIL. As presented in Table 4, AnchorMIL outperformed all baseline models across every metric. Specifically, AnchorMIL achieved the highest AUC (0.885 ± 0.019), accuracy (0.875 ± 0.026), F1-score (0.629 ± 0.042), sensitivity (0.731 ± 0.044), and specificity (0.899 ± 0.036), Our classifier

Table 4: Combined classification (AUC, Accuracy, F1, Sensitivity, Specificity) and regression (MAE, RMSE) performance of various MIL methods on the TCGA-BRCA and OSU datasets. Bold indicates the best classification metric per dataset. Metrics are shown as 95% confidence intervals.

Model	AUC	Accuracy	F1	Sensitivity	Specificity	MAE (%)	RMSE (%)
TCGA-BRCA							
ABMIL	0.823 [0.796,0.850]	0.849 [0.846,0.852]	0.554 [0.542,0.566]	0.629 [0.582,0.676]	0.887 [0.875,0.899]	–	–
MambaMIL	0.824 [0.819,0.829]	0.855 [0.831,0.879]	0.570 [0.556,0.584]	0.638 [0.557,0.720]	0.894 [0.851,0.937]	–	–
MeanMIL	0.819 [0.785,0.853]	0.804 [0.765,0.843]	0.522 [0.506,0.538]	0.717 [0.566,0.868]	0.820 [0.749,0.892]	–	–
S4	0.845 [0.822,0.868]	0.844 [0.807,0.881]	0.568 [0.535,0.601]	0.671 [0.591,0.751]	0.875 [0.819,0.931]	–	–
TransMIL	0.856 [0.843,0.869]	0.846 [0.818,0.874]	0.584 [0.533,0.635]	0.713 [0.694,0.732]	0.869 [0.838,0.901]	–	–
AnchorMIL	**0.885** [0.866,0.904]	**0.875** [0.849,0.902]	**0.629** [0.587,0.671]	**0.731** [0.687,0.776]	**0.899** [0.863,0.935]	5.86 [0.36, 11.36]	7.79 [0.00, 18.86]
OSU							
ABMIL	0.830 [0.826,0.834]	0.767 [0.760,0.774]	0.783 [0.770,0.796]	**0.848** [0.766,0.931]	0.686 [0.593,0.779]	–	–
MambaMIL	0.802 [0.760,0.844]	0.762 [0.738,0.786]	0.726 [0.710,0.742]	0.629 [0.606,0.653]	0.835 [0.768,0.902]	–	–
MeanMIL	0.808 [0.805,0.811]	0.767 [0.753,0.781]	0.769 [0.753,0.786]	0.781 [0.684,0.878]	0.752 [0.635,0.870]	–	–
S4	0.817 [0.802,0.832]	0.752 [0.705,0.799]	0.776 [0.749,0.803]	**0.848** [0.812,0.884]	0.657 [0.534,0.780]	–	–
TransMIL	0.723 [0.691,0.755]	0.629 [0.538,0.721]	0.690 [0.663,0.717]	0.819 [0.678,0.960]	0.438 [0.125,0.752]	–	–
AnchorMIL	**0.858** [0.839,0.877]	**0.800** [0.788,0.812]	**0.790** [0.782,0.798]	0.752 [0.716,0.788]	**0.848** [0.794,0.902]	7.31 [2.18, 12.44]	9.15 [0.00, 18.96]

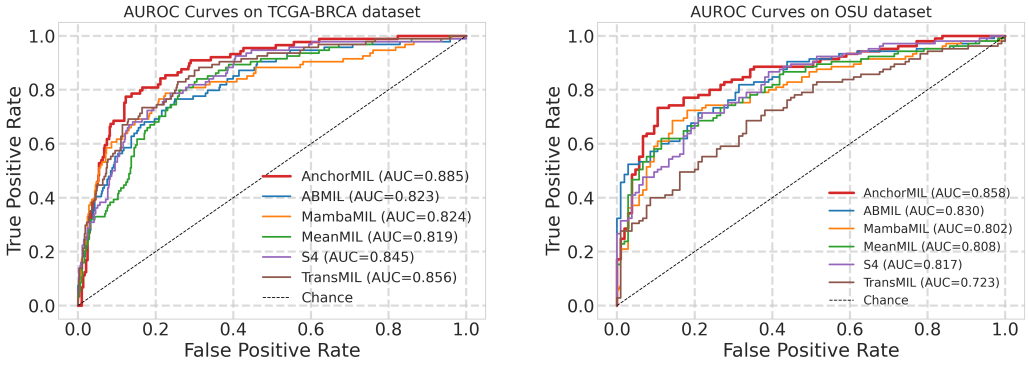
achieved an area under the ROC curve of 0.89 (95% CI 0.87–0.90; $p<0.001$, DeLong's test vs. AUC=0.5). These results suggest that AnchorMIL effectively captures informative patterns within WSIs, contributing to improved predictive performance.

Fig. 4 displays ROC curves generated for different train-test splits of the TCGA dataset, illustrating the consistency of AnchorMIL's predictive performance. Across these splits, AnchorMIL consistently maintained a high AUC, reflecting its stability and reliability in risk prediction tasks. Fig. 5 present the errors associated with the continuous prediction of risk scores.

Overall, these results underscore the effectiveness of our proposed AnchorMIL framework, demonstrating significant enhancements in performance metrics relative to established state-of-the-art methods, as well as robustness and generalizability across diverse datasets.

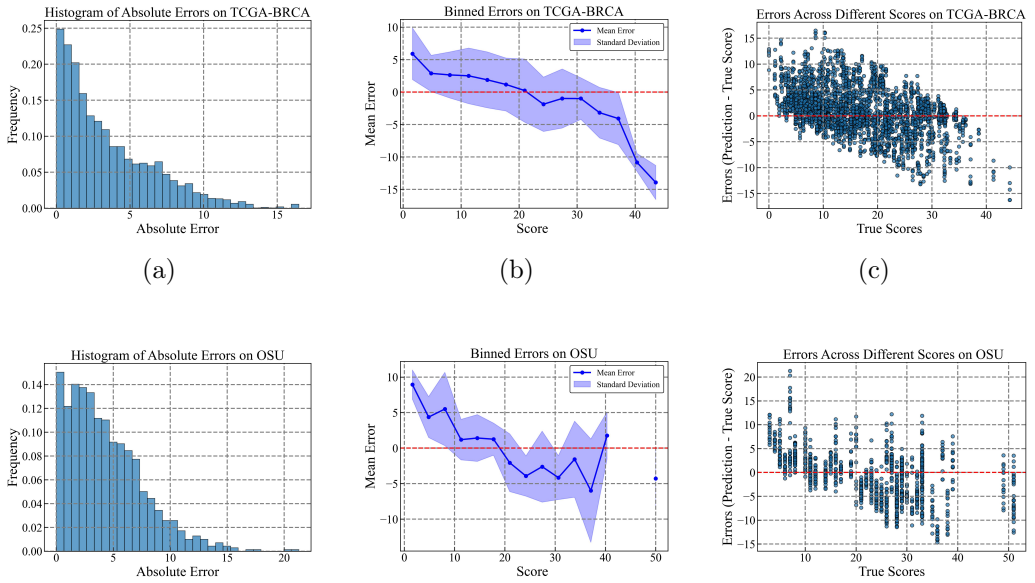
5. Discussion

The proposed AnchorMIL framework offers several key advantages. First, its use of cross-attention and self-attention blocks enables comprehensive feature aggregation, capturing both low-score and high-score patterns from the whole-slide context within a unified Transformer architecture. Second, the anchored regression design with anchors and associated confidence scores combines multiple hypotheses into a single, robust prediction that is resilient to outliers and tissue heterogeneity. Finally, by jointly optimizing regression



(a) (b)

Figure 4: ROC curves for splits of TCGA-BRCA and OSU datasets. The area under the ROC curve (AUC) is reported for each dataset to provide a quantitative measure of classification performance.



(a) (b) (c)

(d) (e) (f) Image 6

Figure 5: Regression error analysis over five independent runs and three data splits for the TCGA-BRCA and OSU cohorts. Panels (a) and (d) show histograms of absolute prediction errors; panels (b) and (c) present binned absolute error distributions; and panels (c) and (f) depict error magnitudes as a function of true score.

and classification objectives, AnchorMIL delivers both precise continuous score estimates and reliable binary risk stratification, making it adaptable to clinical decision-making scenarios.

The anchored regression component is agnostic to the biomarker's biological nature; it simply requires a continuous target range. The parallel classification head only needs a clinically meaningful cutoff point (e.g., positive vs. negative, high-risk vs. low-risk) to provide regulatory guidance to the regression task. Therefore, this framework can be readily adapted to predict other critical continuous biomarkers, such as treatment response scores, protein expression levels, or various prognostic indices. The synergy between the anchored regression for precision and the binary classification for clinical actionability remains constant. This adaptability positions AnchorMIL as a versatile and powerful computational tool for a wide array of problems in digital pathology and beyond.

The experimental results suggest that the proposed **AnchorMIL** framework improves AUC from 0.83 to 0.89 ($\Delta = +7\%$) compared to existing MIL-based methods. One key strength lies in its feature aggregation: incorporating cross-attention and self-attention blocks, which enables the network to capture both local patch-level details and global contextual information from whole slide images (WSIs). This design is particularly effective for handling the spatial heterogeneity inherent in histopathological images.

Another notable aspect is the anchored regression scheme, which employs fixed anchors with learned offsets and confidence scores to provide multiple hypotheses for risk prediction. This mechanism not only enhances the robustness of the predictions but also adds a degree of interpretability, as it allows for an analysis of which anchor best corresponds to the slide-level risk. The framework also benefits from the joint optimization of regression and binary classification losses. This balanced approach allows for both nuanced continuous risk estimation and clear discrete risk stratification, a combination that is especially valuable in clinical settings where both detailed risk assessments and binary treatment decisions are required.

In Fig. 6, we visualize the attention maps from our proposed model on a high-risk WSI, providing insight into the histopathological basis for its risk stratification. The model's attention mechanism appears to identify and assign high importance to complex prognostic features, a process that aligns with the visual evidence used by pathologists in their diagnostic reasoning. For instance, the patches outlined in yellow and blue highlight regions of a high-grade invasive carcinoma that presents a classic prognostic paradox. The

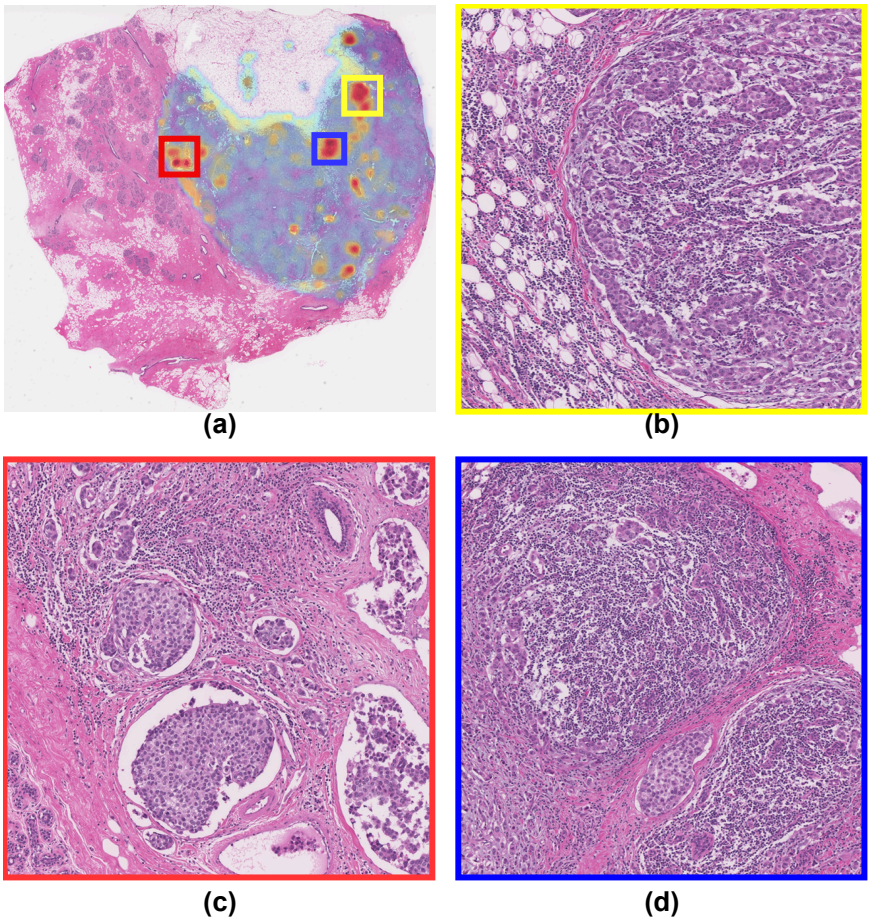


Figure 6: (a) High-risk WSI overlaid with attention heatmap; (b), (c), and (d) inset zoom-boxes highlight the regions receiving the highest attention scores for high risk prediction.

model appropriately highlights features that may be associated with more aggressive tumor biology and potentially poorer prognosis, such as a solid, syncytial growth pattern, notable nuclear pleomorphism, elevated mitotic activity, and evidence of adipose tissue invasion. Concurrently, the model assigns high attention to the brisk infiltrate of tumor-infiltrating lymphocytes (TILs), a strong favorable prognostic factor, particularly in triple-negative or HER2-positive subtypes, which often predicts therapeutic benefit. This demonstrates the model's capacity to recognize the full context of the tumor microenvironment, where aggressive tumor morphology is being counteracted by a robust host immune response. Furthermore, the model shows its ability to prioritize dominant prognostic factors in decisively high-risk cases. The magnified patch outlined in red also features high-grade morphology and a notable TILs infiltrate, but critically, it contains extensive lymphovascular invasion (LVI). The model's high attention to these tumor emboli within vascular spaces reflects the recognition of LVI as a powerful, independent indicator of high metastatic potential that often outweighs the favorable implications of TILs. By correctly classifying this case as high-risk, the model proves it can appropriately weigh the profound negative impact of LVI.

In Fig. 7 most common high-risk patterns are shown: Comodonecrosis and LVI. Comedonecrosis and lymphovascular invasion are powerful morphological indicators of high-risk breast cancer because they directly reflect the aggressive biology measured by a high Oncotype DX score. Comedonecrosis, characterized by central necrosis within ducts packed with high-grade cells, is a visual consequence of runaway proliferation, corresponding directly to the high expression of the heavily weighted proliferation gene group in the ODX panel. Similarly, LVI provides definitive evidence of the tumor's ability to invade and metastasize, a process driven by invasion-related genes that are also quantified by the ODX assay. Therefore, these features are not merely abstract markers of a poor prognosis, they are the physical manifestations of the specific high-risk gene expression signatures that the Oncotype DX test is designed to detect, resulting in a strong and predictable correlation with higher recurrence scores.

Fig. 8 exemplifies the critical discordance that can exist between histological grade and genomic risk assessment in breast cancer. Although the tumor displays aggressive morphological features, such as poor differentiation and high-grade nuclear pleomorphism, its low-risk Oncotype DX score indicates a more favorable prognosis. This discrepancy is likely explained by the absence of other key high-risk features across the whole slide image. Ultimately,

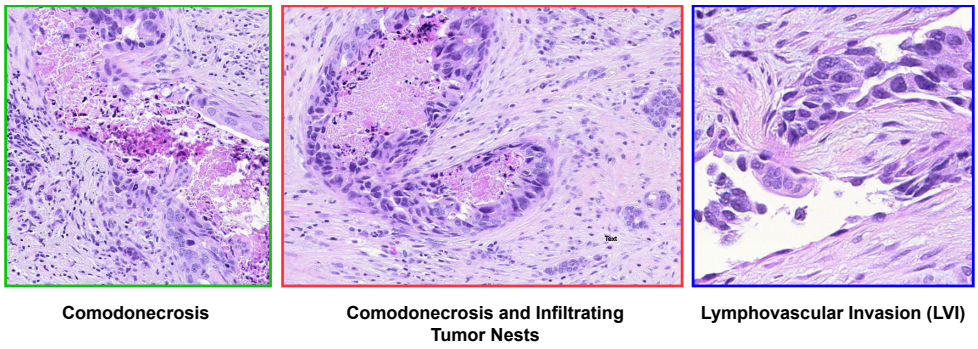


Figure 7: High-risk region of interests. AnchorMIL focuses on Comodonecrosis and Lymphovascular Invasion (LVI) which indicates the model has learned to prioritize features with profound prognostic weight: comedonecrosis as a sign of rapid proliferation inherent to high-grade disease, and LVI as a direct signifier of the tumor's capacity for systemic spread.

this highlights that while histology provides a crucial visual assessment, the tumor's underlying biology, as measured by its gene expression profile, may reveal a lower intrinsic rate of proliferation than the morphology suggests. This underscores the power of genomic testing to refine risk stratification beyond traditional grading, identifying patients with histologically high-grade tumors who can be spared the toxicity of unnecessary adjuvant chemotherapy.

The diversity of datasets used in this study further validates the generalizability of AnchorMIL. Experiments conducted on both the TCGA-BRCA dataset and the OSU in-house dataset indicate that the model performs well across different imaging conditions and patient populations. In particular, the OSU dataset demonstrates the framework's robustness under real-world clinical variability. Performance metrics, as reported in Table 4, show that AnchorMIL outperforms competitive baselines such as ABMIL, MambaMIL, and TransMIL in terms of AUC, Accuracy, F1, Sensitivity, and Specificity. This strong performance underscores the effectiveness of combining anchored regression with advanced attention mechanisms in achieving enhanced risk stratification.

Despite these promising results, certain limitations require further discussion. The efficacy of the attention mechanisms may be affected by the quality and quantity of the available annotations, which can vary across datasets. Future work will explore the integration of additional modalities to further refine the performance and applicability of the proposed framework.

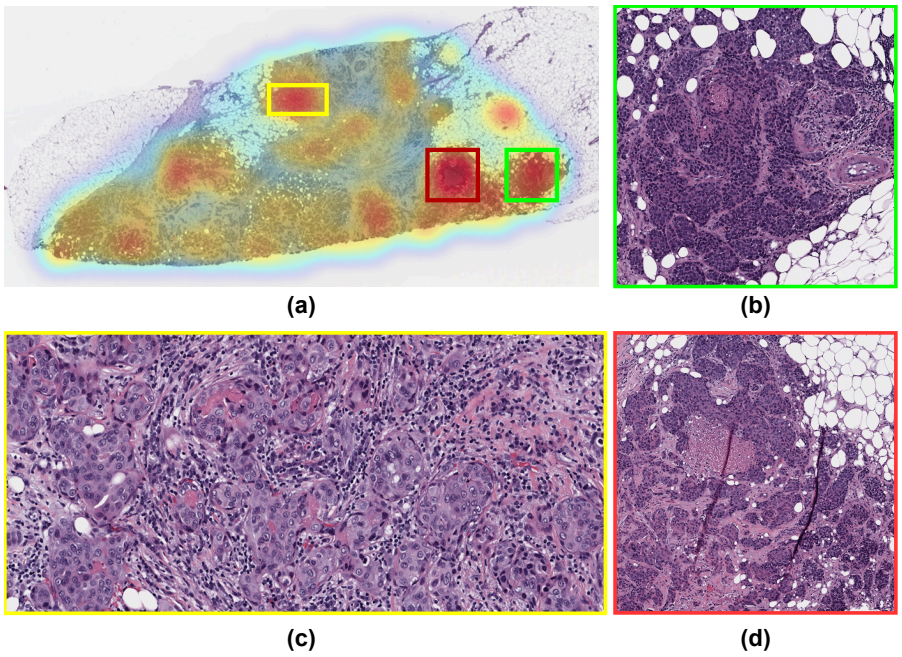


Figure 8: (a) Grade-3, Low ODX scored WSI overlaid with attention heatmap, (b),(c) and (d) inset zoom-boxes highlight the regions receiving the highest attention scores for low risk prediction.

Beyond its predictive performance, the AnchorMIL approach presents several potential practical advantages that could enhance its clinical utility. Unlike conventional genomic assays, which require sending tissue samples to external laboratories and are associated with high cost, long turnaround times, and tissue consumption, AnchorMIL operates directly on routinely available H&E slides on site. This enables rapid and cost-effective risk assessment while preserving tissue for future analyses. Moreover, the framework provides interpretable insights into histopathologic features, supporting more informed clinical decision-making and potentially reducing reliance on expensive molecular testing. Collectively, these attributes highlight AnchorMIL's potential to streamline workflows, improve accessibility, and complement existing risk stratification strategies in diverse clinical settings.

6. Conclusion

In this study, we present AnchorMIL, a novel deep learning framework designed to predict Oncotype DX risk scores directly from H&E-stained whole slide images by integrating a Transformer-based backbone with an anchored regression-classification scheme. AnchorMIL leverages hierarchical attention mechanisms to aggregate patch-level features and employs a confidence-weighted anchoring strategy to enhance prediction robustness and interpretability. Evaluated on 516 TCGA-BRCA slides and 398 institutional slides, AnchorMIL demonstrated a substantial improvement in high-vs-low risk classification, increasing the AUC from 0.83 to 0.89 compared to existing MIL baselines, while maintaining a balanced sensitivity–specificity profile. When applied at the clinically accepted cutoff of 25, the model could have potentially obviated approximately 40% of genomic assays in our cohorts, thereby enabling faster treatment decisions and reducing healthcare costs.

Despite these promising results, the study has several limitations. It remains retrospective in design, is limited to data from U.S.-based academic scanners, and lacks racial diversity proportional to national breast cancer incidence. Additionally, while anchor attention heatmaps offer insights into model interpretability, their biological plausibility requires validation through expert pathology consensus. The efficacy of the attention mechanisms may also be influenced by the quality and quantity of available annotations, which can vary across datasets. Future work will explore the integration of additional modalities to further refine the performance and applicability of the proposed framework.

From a clinical perspective, AnchorMIL offers a compelling tool for augmenting breast cancer risk stratification, particularly in hormone-receptor-positive early-stage disease. By providing rapid, cost-effective, and interpretable predictions directly from routinely available histology, the model has the potential to streamline clinical workflows and reduce reliance on expensive genomic assays. This could be especially impactful in resource-limited settings or institutions with delayed access to molecular testing. Furthermore, the model's ability to highlight biologically meaningful features such as lymphovascular invasion and comedonecrosis aligns with established prognostic markers, reinforcing its clinical relevance. Ultimately, AnchorMIL supports a more personalized approach to breast cancer management, helping to identify patients who may safely forgo chemotherapy and those who may benefit from more aggressive treatment, thereby improving outcomes while minimizing overtreatment.

References

[1] Rebecca L Siegel, Tyler B Kratzer, Angela N Giaquinto, Hyuna Sung, and Ahmedin Jemal. Cancer statistics, 2025. *Ca*, 75(1):10, 2025.

[2] Takeo Fujii, Fanny Le Du, Lianchun Xiao, Takahiro Kogawa, Carlos H Barcenas, Ricardo H Alvarez, Vicente Valero, Yu Shen, and Naoto T Ueno. Effectiveness of an adjuvant chemotherapy regimen for early-stage breast cancer: a systematic review and network meta-analysis. *JAMA oncology*, 1(9):1311–1318, 2015.

[3] Bonnie N Joe, Harold J Burstein, and Sadhna R Vora. Clinical features, diagnosis, and staging of newly diagnosed breast cancer. *UpToDate. Burstein H, Vora SR (eds.). Waltham, MA: UpToDate*, 2019.

[4] Nadia Howlader, Sean F Altekruse, Christopher I Li, Vivien W Chen, Christina A Clarke, Lynn AG Ries, and Kathleen A Cronin. Us incidence of breast cancer subtypes defined by joint hormone receptor and her2 status. *Journal of the National Cancer Institute*, 106(5):dju055, 2014.

[5] Shuyi Chen, Christopher Thacker, Shengxuan Wang, Katelyn A Young, Rebecca L Hoffman, and Joseph A Blansfield. Adherence disparities and utilization trends of oncotype dx assay: a national cancer database study. *Journal of Surgical Research*, 286:65–73, 2023.

[6] Ran Song, Dong-Eun Lee, Eun-Gyeong Lee, Seeyoun Lee, Han-Sung Kang, Jai Hong Han, Keun Seok Lee, Sung Hoon Sim, Heejung Chae, Youngmee Kwon, et al. Clinicopathological factors associated with oncotype dx risk group in patients with er+/her2-breast cancer. *Cancers*, 15(18):4451, 2023.

[7] Joseph A Sparano, Robert J Gray, Della F Makower, Kathleen I Pritchard, Kathy S Albain, Daniel F Hayes, Charles E Geyer Jr, Elizabeth C Dees, Matthew P Goetz, John A Olson Jr, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 379(2):111–121, 2018.

[8] Joseph A Sparano. Tailorx: trial assigning individualized options for treatment (rx). *Clinical breast cancer*, 7(4):347–350, 2006.

[9] M Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N Gurcan. Computerized classification of intraductal breast lesions using histopathological images. *IEEE Transactions on Biomedical Engineering*, 58(7):1977–1984, 2011.

[10] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019.

[11] Mingxi Ouyang, Yuqiu Fu, Renao Yan, ShanShan Shi, Xitong Ling, Lianghui Zhu, Yonghong He, and Tian Guan. Mergeup-augmented semi-weakly supervised learning for wsi classification. *arXiv preprint arXiv:2408.12825*, 2024.

[12] Ziyu Su, Amanda Rosen, Robert Wesolowski, Gary Tozbikian, M Khalid Khan Niazi, and Metin N Gurcan. Deep-odx: An efficient deep learning tool to risk stratify breast cancer patients from histopathology images. In *Medical Imaging 2024: Digital and Computational Pathology*, volume 12933, pages 34–39. SPIE, 2024.

[13] Haojia Li, Jon Whitney, Kaustav Bera, Hannah Gilmore, Mangesh A Thorat, Sunil Badve, and Anant Madabhushi. Quantitative nuclear histomorphometric features are predictive of oncotype dx risk categories in ductal carcinoma in situ: preliminary findings. *Breast cancer research*, 21:1–16, 2019.

[14] Jon Whitney, German Corredor, Andrew Janowczyk, Shridar Ganesan, Scott Doyle, John Tomaszewski, Michael Feldman, Hannah Gilmore, and Anant Madabhushi. Quantitative nuclear histomorphometry predicts oncotype dx risk categories for early stage er+ breast cancer. *BMC cancer*, 18:1–15, 2018.

[15] Jialiang Yang, Jie Ju, Lei Guo, Binbin Ji, Shufang Shi, Zixuan Yang, Songlin Gao, Xu Yuan, Geng Tian, Yuebin Liang, et al. Prediction of her2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Computational and structural biotechnology journal*, 20:333–342, 2022.

[16] Kevin M Boehm, Omar SM El Nahhas, Antonio Marra, Michele Waters, Justin Jee, Lior Braunstein, Nikolaus Schultz, Pier Selenica, Hannah Y Wen, Britta Weigelt, et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *Nature Communications*, 16(1):2106, 2025.

[17] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern recognition*, 77:329–353, 2018.

[18] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[19] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[20] Christopher J Ricketts, Aguirre A De Cubas, Huihui Fan, Christof C Smith, Martin Lang, Ed Reznik, Reanne Bowlby, Ewan A Gibb, Rehan Akbani, Rameen Beroukhim, et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell reports*, 23(1):313–326, 2018.

[21] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple

instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[22] Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 296–306. Springer, 2024.

[23] Ziyu Su, Mostafa Rezapour, Usama Sajjad, Shuo Niu, Metin Nafi Gurcan, and Muhammad Khalid Khan Niazi. Cross-attention-based saliency inference for predicting cancer metastasis on whole slide images. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[24] Dawid Rymarczyk, Adam Pardyl, Jarosław Kraus, Aneta Kaczyńska, Marek Skomorowski, and Bartosz Zieliński. Protomil: Multiple instance learning with prototypical parts for whole-slide image classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 421–436. Springer, 2022.

[25] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18802–18812, 2022.

[26] Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79:102464, 2022.

[27] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob Van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[28] Gianpaolo Bontempo, Angelo Porrello, Federico Bolelli, Simone Calderara, and Elisa Ficarra. Das-mil: distilling across scales for mil classification of histological wsis. In *International conference on medical image*

computing and computer-assisted intervention, pages 248–258. Springer, 2023.

[29] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical image analysis*, 88:102802, 2023.

[30] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[31] 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*, pages 770–778, 2016.

[32] Ziyu Su, Muhammad Khalid Khan Niazi, Thomas E Tavolara, Shuo Niu, Gary H Tozbikian, Robert Wesolowski, and Metin N Gurcan. Bcr-net: A deep learning framework to predict breast cancer recurrence from histopathology images. *Plos one*, 18(4):e0283562, 2023.

[33] 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.

[34] Frederick M Howard, James Dolezal, Sara Kochanny, Galina Khramtsova, Jasmine Vickery, Andrew Srisuwananukorn, Anna Woodard, Nan Chen, Rita Nanda, Charles M Perou, et al. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *NPJ Breast Cancer*, 9(1):25, 2023.