Double-tier Attention based Multi-label Learning Network for Predicting Biomarkers from Whole Slide Images of Breast Cancer

Mingkang Wang[1,2], Tong Wang[1], Fengyu Cong[1,2], Cheng Lu[3](✉), and Hongming Xu[1,2](✉)

[1] School of Biomedical Engineering, Faulty of Medicine, Dalian University of Technology, Dalian 116024, China
mxu@dlut.edu.cn
[2] Key Laboratory of Integrated Circuit and Biomedical Electronic System, Liaoning Province, Dalian University of Technology, Dalian 116024, China
[3] Department of Radiology, Guangdong Provincial People's Hospital, Southern Medical University, Guangzhou 510080, China
lucheng@gdph.org.cn

Abstract. Hematoxylin and eosin (H&E) staining offers the advantages of low cost and high stability, effectively revealing the morphological structure of the nucleus and tissue. Predicting the expression levels of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) from H&E stained slides is crucial for reducing the detection cost of the immunohistochemistry (IHC) method and tailoring the treatment of breast cancer patients. However, this task faces significant challenges due to the scarcity of large-scale and well-annotated datasets. In this paper, we propose a double-tier attention based multi-label learning network, termed as DAMLN, for simultaneous prediction of ER, PR, and HER2 from H&E stained WSIs. Our DAMLN considers slides and their tissue tiles as bags and instances under a multiple instance learning (MIL) setting. First, the instances are encoded via a pretrained CTransPath model and randomly divided into a set of pseudo bags. Pseudo-bag guided learning via cascading the multi-head self-attention (MSA) and linear MSA blocks is then conducted to generate pseudo-bag level representations. Finally, attention-pooling is applied to class tokens of pseudo bags to generate multiple biomarker predictions. Our experiments conducted on large-scale datasets with over 3000 patients demonstrate great improvements over comparative MIL models. The code is available at https://github.com/PerrySkywalker/DAMLN.

Keywords: Brest cancer · Biomarker prediction · Attention mechanism.

1 Introduction

The molecular classification is pivotal for guiding the diagnosis and treatment of breast cancer [19]. Identifying the expression levels of key predictive biomarkers such as ER, PR, and HER2 through IHC staining is considered the gold standard for diagnosing and staging breast cancer [21]. Currently, biomarker status is typically determined using separate IHC staining for each biomarker [16]. However, IHC staining is costly, time-consuming, and often missing in clinical practice due to resource constraints, hindering its widespread application [10,9]. In contrast, H&E staining is more efficient and cost-effective, providing morphological insights into tissues and cells, which can also inform IHC biomarker status. Therefore, there is a pressing need to develop effective AI models for predicting IHC biomarkers using H&E stained whole slide images (WSIs).

Recent studies have explored deep learning models for predicting IHC biomarker status using H&E stained histological images [8]. For example, Couture et al. [2] utilized a pretrained VGG16 network to extract features from H&E stained tissue microarrays (TMA), and then predicted the ER status of breast cancer patients via support vector machine (SVM). Rawat et al. [17] trained ResNet models to extract "tissue fingerprints" from H&E stained TMA images of breast cancer patients. Subsequently, they fed patch-level "fingerprints" into three neural networks to predict the ER, PR, and HER2 status, respectively. Gamble et al. [4] trained three fully supervised InceptionV3 models, using H&E stained histological images as inputs, to make patch-level predictions for ER, PR, and HER2 status. They then quantified patch-level predictions and generated WSI-level predictions using a regularized regression model. Although these approaches have reported promising performances in IHC biomarker prediction, they all rely on patch-level classification models, assuming that all histological image patches inherit patient-level biomarker labels. However, this assumption overlooks the heterogeneity of the tumor-immune microenvironment, resulting in noisy training of the biomarker prediction model [6,24].

To address the challenges of lacking patch-level biomarker labels for WSIs, MIL emerges as a promising approach, requiring only bag-level labels for model training. Recently, embedded-space (ES) based MIL methods have gained traction among researchers for addressing WSI classification tasks. In the ES paradigm, a pretrained model is leveraged for feature extraction, embedding all instances as low-dimensional feature vectors. Then, a bag-level aggregator, often based on attention-pooling or MSA in Transformer [20], is employed to aggregate features and derive final bag-level representations. The popular ES-based MIL models for WSI classifications include attention-based MIL (ABMIL) [7], clustering-constrained attention MIL (CLAM) [12], Transformer based MIL (TransMIL) [18], and double-tier feature distillation MIL (DTFD) [23]. Using the ABMIL framework, Naik et al. [14] devised a weakly-supervised deep neural network to identify ER status from H&E-stained WSIs. Training deep MIL models necessitates a large quantity of WSIs with diverse data; otherwise, effective optimization of deep MIL models is impeded, resulting in poor classification performance. However, amassing a large WSI cohort is financially and logistically challenging due to data scarcity and collection difficulties. Furthermore, there are multiple IHC biomarkers such as ER, PR, and HER2, indicating a need for multi-label learning integrated with MIL to better fulfill clinical applications.

In this paper, we propose a multi-label learning model called DAMLN for simultaneous prediction of multiple IHC biomarkers using H&E stained WSIs.

The contributions of this paper are: (1) We develop a pseudo-bag guided learning approach that enhances the diversity and quantity of bags for effectively training the MIL framework, thereby improving prediction performance. This approach decomposes a WSI into several randomly generated pseudo-bags, reducing the computational burden of MIL by decreasing bag size. (2) By stacking standard MSA and linear MSA blocks, our DAMLN model can better learn global interactions among instances when generating pseudo-bag representations, resulting in improved aggregation of instance-level embeddings compared to other MIL models. (3) To the best of our knowledge, our DAMLN is the first multi-label learning model that exploits the correlation among biomarkers to enhance accuracy and efficiency of simultaneously predicting ER, PR, and HER2 status via H&E stained WSIs.

2 Methods

Figure 1 shows the overview of our DAMLN model which includes three modules: pseudo-bag generation, pseudo-bag guided learning, and multi-label learning prediction. The details of our DAMLN model are described as follows.

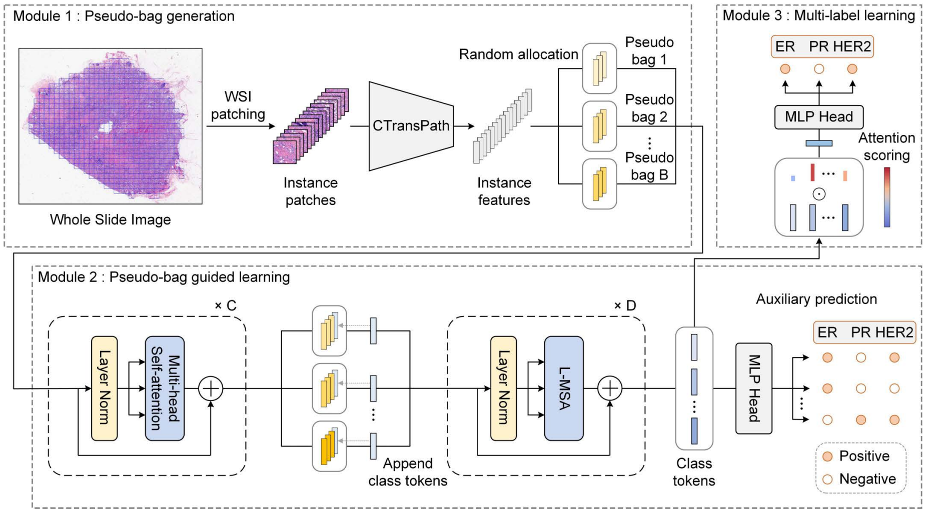

Fig. 1. Overview of our DAMLN model. Pseudo-bags are generated from patch-level embedding, and pseudo-bag guided learning is conducted to assist in learning effective pseudo-bag representations. Multi-label learning is then performed to predict multiple IHC biomarkers.

2.1 Problem Formulation

Given a dataset $\mathcal{W} = \{W_1, W_2, ..., W_N\}$ consisting of N WSIs, each WSI W_i has multiple labels $\{y_{i,1}, y_{i,2}, ..., y_{i,M}\}$, where $y_{i,j} \in \{0, 1\}, j = \{1, 2, ..., M\}$ represents different biomarker labels (e.g., ER±, PR±, HER2±). We then break down each WSI W_i into numerous patches $\{p_{i,1}, p_{i,2}, ..., p_{i,n_i}\}$, where n_i is the number of patches obtained from the i-th WSI; $p_{i,n_i} \in \mathbb{R}^{W \times H \times 3}$, where W and H represent the width and height of the patch. In the MIL paradigm, all patches $\{p_{i,1}, p_{i,2}, ..., p_{i,n_i}\}$ from W_i constitute a bag, and each patch within a bag is treated as an instance. The bag has the same labels as the corresponding WSI, while the labels for its instances are unknown. Our objective is to effectively aggregate instance-level embeddings to generate comprehensive bag-level representations for the simultaneous prediction of multiple IHC biomarkers.

2.2 Pseudo-bag Generation

In WSIs, background regions (i.e., nearly white colors) are irrelevant for biomarker prediction and only introduce noise and computational overhead, which are excluded for analysis by thresholding [12]. The foreground regions containing different tissue components are then divided into a set of non-overlapping image patches at 20× magnification (224×224 pixels per patch). We employ the CTransPath [22] as a feature extractor to derive features from tiled patches. CTransPath, a hybrid model that combines a convolutional neural network (CNN) with a multi-scale Swin Transformer [11], has been pretrained on approximately 15 million histological image patches, making it a potent feature extractor. Let us denote the embedded feature vectors corresponding to patches $\{p_{i,1}, p_{i,2}, ..., p_{i,n_i}\}$ as $\{f_{i,1}, f_{i,2}, ..., f_{i,n_i}\}$, where $f_{i,n_i} \in \mathbb{R}^l$. After patch embedding, we randomly partition instance features of each bag into B roughly equal subsets called pseudo-bags, denoted as $H = \{h_1, h_2, ...h_B\}$, where $h_j \in \mathbb{R}^{n \times l}$, $1 \leq j \leq B$, and n represents the size of pseudo-bag and may slightly vary across different pseudo-bags. Each pseudo-bag inherits the label of its parent bag. This generation of pseudo-bags can effectively address the challenge of insufficient WSIs for training MIL models, and enhance the generalization ability and classification performance.

2.3 Pseudo-bag Guided Learning

As shown in Figure 1, the generated pseudo-bags are fed into a sequential of C residual-connected MSA blocks to learn long-range dependencies among different instances within each pseudo-bag. The MSA block comprises several parallel standard softmax self-attention operations, i.e.,

$$A\left(Q,\ K,\ V\right) = soft\max\left(\frac{QK^T}{\sqrt{D}}\right)V, \tag{1}$$

where $Q = h_j W_Q$, $K = h_j W_K$, $V = h_j W_V$, and $W_Q, W_K, W_V \in \mathbb{R}^{l \times l}$ are learnable linear projection matrices. The attention matrix A has a computational complexity of $O(2n^2 l)$.

Given the considerable size of WSIs, with n often reaching thousands, the computational burden of stacking numerous standard MSA blocks is substantial. In this study, we opt for 2 standard MSA blocks (C=2) to strike a balance between computational efficiency and classification performance.

The softmax function acts as a similarity measure between Q and K [1], and it can be substituted with a decomposable kernel $\phi(\cdot)$, following the commutative property of multiplication, i.e.,

$$A_l(Q, K, V) = \phi(Q)\left(\phi\left(K^T\right)V\right). \tag{2}$$

Compared to computing A, computing A_l has a significantly smaller computational complexity of $O(2nl^2)$, as n is much larger than l in WSI classifications. By employing A_l as a replacement for A in the similarity measure, we further stack 2 linear MSA blocks (D=2 in Figure 1) to deeply capture the internal correlation among different instance embeddings. In this study, we use the Sigmoid and Tanh functions as kernel functions for Q and K, respectively, i.e.,

$$A_l(Q, K, V) = Sigmoid(Q)\left(Tanh\left(K^T\right)V\right). \tag{3}$$

Note that before feeding instance embeddings into linear MSA blocks, a class token (i.e., x^{class}) is prepended to the beginning of each pseudo-bag. This token serves to encode the class information associated with that pseudo-bag. Finally, all class tokens of pseudo-bags are fed into an auxiliary multi-layer perceptron (MLP) classifier for multi-label classifications regarding ER, PR and HER2 status. The loss function for the auxiliary classification is identical to that of the WSI-level multi-label learning (see Equation (6)).

2.4 Multi-label Learning Prediction

The class tokens (i.e., x_i^{class}) of pseudo-bags are fed into attention-pooling block that aggregates them into the WSI-level representation x_s, which is computed as:

$$x_s = \sum_{i}^{B} a_i x_i^{class}, \tag{4}$$

$$a_i = \frac{\exp\left\{\left(\tanh\left(x_i^{class}W_Z\right) \odot sigmoid\left(x_i^{class}W_U\right)\right)W_{ZU}\right\}}{\sum_{j=1}^{B}\exp\left\{\left(\tanh\left(x_j^{class}W_Z\right) \odot sigmoid\left(x_j^{class}W_U\right)\right)W_{ZU}\right\}}, \tag{5}$$

where $W_Z, W_U \in \mathbb{R}^{l \times d}$ and $W_{ZU} \in \mathbb{R}^{d \times 1}$ are learnable linear projection matrices, and $\odot$ is an element-wise multiplication. The WSI-level representation is connected with an MLP classifier for the multi-label prediction. The overall loss function L_i based on binary cross entropy is computed as:

$$L_i = mean\left(\sum_{j=1}^{M} -w_{i,j}\left(y_{i,j}\log\left(\hat{y_{i,j}}\right) + \left(1 - y_{i,j}\right)\log\left(1 - \hat{y_{i,j}}\right)\right)\right), \tag{6}$$

where $y_{i,j}$ and $\hat{y_{i,j}}$ represent ground truths and predictions, respectively. The problem of missing certain labels usually occurs in multi-label learning, restricting model optimization to only samples with biomarker labels. The parameter $w_{i,j}$ facilitates binary cross-entropy computation solely on samples with biomarker labels. Specifically, if y_c exists, then w_c is set to 1; otherwise, it is set to 0. The function $mean(\cdot)$ represents an adaptive averaging operation applied based on the available biomarker labels.

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

QHSU Dataset. QHSU dataset comprises 2384 H&E stained WSIs of breast cancer patients, each representing one patient. These WSIs were scanned under 40× magnification (0.2511 um/pixel) at Qilu Hospital of Shandong University in China. Biomarker labels for H&E stained WSIs are derived from diagnostic reports, where experienced pathologists diagnosed biomarker status by assessing the corresponding IHC stained slides. Notably, the number of patients with HER2 labels (1688 patients) is fewer than those of ER and PR labels (2384 patients).

TCGA-BRCA Dataset. TCGA-BRCA is a publicly available dataset comprising breast cancer WSIs collected by The Cancer Genome Atlas (TCGA) project. Poor-quality WSIs exhibiting severe artifacts or lacking IHC biomarker labels are filtered out. As a result, we collect 757 WSIs with ER and PR labels, and 745 WSIs with HER2 labels, forming the TCGA-BRCA cohort for independent testing.

Table 1. Distribution of IHC biomarker labels in QHSU and TCGA-BRCA datasets

Datasets	QHSU			TCGA-BRCA		
Labels	ER	PR	HER2	ER	PR	HER2
Positive	1822	1750	470	583	509	111
Negative	562	634	1218	174	248	634

Table 1 summarizes the data distribution of all patients across our QHSU and TCGA-BRCA datasets. Using the QHSU dataset, we first conduct ablation experiments to assess the empirical settings in our DAMLN model. Subsequently, we perform 5-fold cross-validation and compute the average results as internal testing. Finally, models saved from the 5-fold cross-validation on the QHSU dataset are used for external testing on the TCGA-BRCA dataset, and the average results are reported. Our DAMLN model and other comparative models mentioned in this study were implemented using Python and PyTorch [15], and trained on an NVIDIA GeForce RTX4090 GPU. We employed the AdamW as the optimizer, with a learning rate of 1e-4 and weight decay of 1e-5. Given the variability in size among WSIs, we set the batch size to 1. Within each batch, the auxiliary prediction followed by WSI-level prediction is sequentially performed to optimize our DAMLN model. The model performance is evaluated using accuracy (ACC) and area under the receiver operating characteristic curve (AUC).

3.2 Results and Discussion

Ablation Study. We conducted an ablation study on DAMLN using the QHSU Dataset to examine the impact of pseudo-bag quantity B, patch embedding encoder, and MSA block settings. Table 2 lists our ablation study results. As outlined in Table 2, we increased the value of B from 1 to 7 with a step of 2. Considering the training time and AUC values, we found that setting B to 5 yields the most promising performance. Subsequently, we replaced all linear MSA blocks with standard MSA, resulting in a decline in ER and PR AUC values alongside an increase in training time. Finally, when we substituted the CTransPath patch encoder with a pre-trained ResNet50 [5] on the ImageNet dataset [3], nearly all performance metrics exhibit a marked decrease, with AUC values experiencing a notable reduction of 0.37%-2.31%. This highlights the efficacy of patch-level embedding using the CTransPath model pretrained specifically for histological images.

Table 2. Ablation experiments in terms of the value of B, settings on MSA blocks, and patch-level encoder.

Models	No. of	Time/	ER		PR		HER2	
	Bags	Epoch	AUC	ACC	AUC	ACC	AUC	ACC
DAMLN	B=1	189s	0.9129	0.8393	0.8516	0.7928	0.8750	0.7986
DAMLN	B=3	91s	0.9141	0.8545	0.8633	0.8184	**0.8920**	**0.8281**
DAMLN	B=5	75s	**0.9232**	0.8528	**0.8678**	0.8167	0.8884	0.8044
DAMLN	B=7	60s	0.9178	**0.8603**	0.8616	**0.8217**	0.8885	0.8270
Standard MSA	B=5	117s	0.9133	0.8565	0.8568	0.8171	0.8866	0.8257
ResNet50	B=5	80s	0.9029	0.8334	0.8447	0.7898	0.8847	0.8166

Internal Comparison. We compare our DAMLN model with the following state-of-the-art (SOTA) MIL models: ABMIL [7], Single-attention-branch CLAM-SB [12], Multi-attention-branch CLAM-MB [12], TransMIL [18], and DTFD [23]. It is worth noting that all compared methods were implemented in a single label learning approach and adopted CTransPath as patch-level encoder. Table 3 list comparisons in terms of ER, PR, and HER2 predictions on the QHSU dataset. As observed in Table 3, our DAMLN model outperforms all the compared MIL models across different biomarker predictions. Specifically, our DAMLN model achieves the highest AUC values of 0.9232, 0.8678, and 0.8884 for ER, PR, and HER2, respectively, as well as the highest ACC value of 0.8167 for PR. Although our DAMLN achieves slightly lower ACC values on ER and

HER2 predictions compared to some other MIL models (e.g., CLAM), it is important to consider the severe label imbalance. In such cases, AUC is considered a more comprehensive measure of performance. Unlike existing MIL models that require separate training processes for each biomarker, our DAMLN model offers the advantage of requiring a single-pass training to predict ER, PR, and HER2 status concurrently.

Table 3. Comparisons with SOTA MIL methods on the QHSU dataset.

Models	ER		PR		HER2	
	AUC	ACC	AUC	ACC	AUC	ACC
ABMIL [7]	0.9021	0.8477	0.8523	0.8071	0.8823	0.8198
CLAM-SB [12]	0.9013	0.8507	0.8453	0.8028	0.8806	**0.8276**
CLAM-MB [12]	0.8997	**0.8574**	0.8483	0.8058	0.8856	0.8240
TransMIL [18]	0.9029	0.8355	0.8525	0.7878	0.8797	0.8129
DTFD [23]	0.8930	0.8532	0.8377	0.8100	0.8758	0.8270
Ours (DAMLN)	**0.9232**	0.8528	**0.8678**	**0.8167**	**0.8884**	0.8044

External Comparison. Table 4 lists independent comparisons on TCGA-BRCA dataset in terms of ER, PR, and HER2 predictions. As observed in Table 4, our DAMLN model provides significant improvements compared with recent biomarker prediction studies on TCGA-BRCA dataset. Particularly, our model has over 3% and 6% improvements in ER and PR predictions compared to these relevant studies. In addition, our model outperforms other SOTA MIL models in the independent testing dataset, indicating the efficacy of pseudo-bag guided learning and double-tier attentions developed in this study.

Table 4. External comparisons with relevant studies and SOTA MIL models on TCGA-BRCA dataset.

Models	ER		PR		HER2	
	AUC	ACC	AUC	ACC	AUC	ACC
Gamble et al. [4]	0.83	-	0.72	-	0.58	-
Kather et al. [8]	0.82	-	0.74	-	-	-
Naik et al. [14]	0.85	-	-	-	-	-
Lu et al. [13]	-	-	-	-	0.75	-
ABMIL [7]	0.8646	0.8421	0.8026	0.7604	0.7274	0.7791
CLAM-SB [12]	0.8727	0.8492	0.7932	0.7540	**0.7522**	0.7632
CLAM-MB [12]	0.8721	**0.8515**	0.7928	0.7614	0.7487	0.7710
TransMIL [18]	0.8653	0.8403	0.7906	0.7635	0.7224	0.7603
DTFD [23]	0.8686	0.8418	0.7817	0.7543	0.7374	0.7036
Ours (DAMLN)	**0.8801**	0.8476	**0.8090**	**0.7707**	0.7521	**0.7826**

4 Conclusion

In this paper, we propose a double-tier attention-based multi-label learning network called DAMLN, capable of simultaneously predicting multiple biomarkers from H&E stained WSIs of breast cancer. Our DAMLN model integrates standard MSA and linear MSA blocks to aggregate instance-level embeddings into comprehensive slide-level representations. Pseudo-bag guided learning and multi-label learning are developed to effectively train MIL models and simultaneously predict multiple IHC biomarkers. Experiments conducted on two large datasets demonstrate the advantages of our model in predicting ER, PR, and HER2 status for breast cancer patients.

Acknowledgments. This work was supported in part by the the National Natural Science Foundation of China (82102135), and the Fundamental Research Funds for Central Universities (DUT23YG130).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Hoffman, J.: Hydra attention: Efficient attention with many heads. In: European Conference on Computer Vision. pp. 35–49. Springer (2022)
2. Couture, H.D., Williams, L.A., Geradts, J., Nyante, S.J., Butler, E.N., Marron, J., Perou, C.M., Troester, M.A., Niethammer, M.: Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype. NPJ breast cancer **4**(1), 30 (2018)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
4. Gamble, P., Jaroensri, R., Wang, H., Tan, F., Moran, M., Brown, T., Flament-Auvigne, I., Rakha, E.A., Toss, M., Dabbs, D.J., et al.: Determining breast cancer biomarker status and associated morphological features using deep learning. Communications medicine **1**(1), 14 (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hsieh, W.C., Budiarto, B.R., Wang, Y.F., Lin, C.Y., Gwo, M.C., So, D.K., Tzeng, Y.S., Chen, S.Y.: Spatial multi-omics analyses of the tumor immune microenvironment. Journal of Biomedical Science **29**(1), 96 (2022)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
8. Kather, J.N., Heij, L.R., Grabsch, H.I., Loeffler, C., Echle, A., Muti, H.S., Krause, J., Niehues, J.M., Sommer, K.A., Bankhead, P., et al.: Pan-cancer image-based detection of clinically actionable genetic alterations. Nature cancer **1**(8), 789–799 (2020)

9. Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al.: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature medicine **25**(7), 1054–1056 (2019)
10. Lim, C., Tsao, M., Le, L., Shepherd, F., Feld, R., Burkes, R., Liu, G., Kamel-Reid, S., Hwang, D., Tanguay, J., et al.: Biomarker testing and time to treatment decision in patients with advanced nonsmall-cell lung cancer. Annals of Oncology **26**(7), 1415–1421 (2015)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
12. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
13. Lu, W., Toss, M., Dawood, M., Rakha, E., Rajpoot, N., Minhas, F.: Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. Medical Image Analysis **80**, 102486 (2022)
14. Naik, N., Madani, A., Esteva, A., Keskar, N.S., Press, M.F., Ruderman, D., Agus, D.B., Socher, R.: Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. Nature communications **11**(1), 5727 (2020)
15. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
16. Ramos-Vara, J.A.: Technical aspects of immunohistochemistry. Veterinary pathology **42**(4), 405–426 (2005)
17. Rawat, R.R., Ortega, I., Roy, P., Sha, F., Shibata, D., Ruderman, D., Agus, D.B.: Deep learned tissue "fingerprints" classify breast cancers by er/pr/her2 status from h&e images. Scientific reports **10**(1), 7275 (2020)
18. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
19. Subik, K., Lee, J.F., Baxter, L., Strzepek, T., Costello, D., Crowley, P., Xing, L., Hung, M.C., Bonfiglio, T., Hicks, D.G., et al.: The expression patterns of er, pr, her2, ck5/6, egfr, ki-67 and ar by immunohistochemical analysis in breast cancer cell lines. Breast cancer: basic and clinical research **4**, 117822341000400004 (2010)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
21. Waks, A.G., Winer, E.P.: Breast cancer treatment: a review. Jama **321**(3), 288–300 (2019)
22. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
23. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: 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. pp. 18802–18812 (2022)
24. Zilenaite, D., Rasmusson, A., Augulis, R., Besusparis, J., Laurinaviciene, A., Plancoulaine, B., Ostapenko, V., Laurinavicius, A.: Independent prognostic value of

intratumoral heterogeneity and immune response features by automated digital immunohistochemistry analysis in early hormone receptor-positive breast carcinoma. Frontiers in oncology **10**, 950 (2020)