Point Transformer with Federated Learning for Predicting Breast Cancer HER2 Status from Hematoxylin and Eosin-Stained Whole Slide Images

Bao Li[1, 2*], Zhenyu Liu[2*], Lizhi Shao[2], Bensheng Qiu[1], Hong Bu[3†], Jie Tian[1, 2, 4†]

[1]Center for Biomedical Imaging, University of Science and Technology of China, Hefei, China
[2]CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Department of Pathology, West China Hospital, Sichuan University, Chengdu, China
[4]Key Laboratory of Big Data-Based Precision Medicine, Ministry of Industry and Information Technology, School of Engineering Medicine, Beihang University, Beijing, China
libao1506@mail.ustc.edu.cn, bqiu@ustc.edu.cn, hongbu@scu.edu.cn, {zhenyu.liu, lizhi.shao, jie.tian}@ia.ac.cn

Abstract

Directly predicting human epidermal growth factor receptor 2 (HER2) status from widely available hematoxylin and eosin (HE)-stained whole slide images (WSIs) can reduce technical costs and expedite treatment selection. Accurately predicting HER2 requires large collections of multi-site WSIs. Federated learning enables collaborative training of these WSIs without gigabyte-size WSIs transportation and data privacy concerns. However, federated learning encounters challenges in addressing label imbalance in multi-site WSIs from the real world. Moreover, existing WSI classification methods cannot simultaneously exploit local context information and long-range dependencies in the site-end feature representation of federated learning. To address these issues, we present a point transformer with federated learning for multi-site HER2 status prediction from HE-stained WSIs. Our approach incorporates two novel designs. We propose a dynamic label distribution strategy and an auxiliary classifier, which helps to establish a well-initialized model and mitigate label distribution variations across sites. Additionally, we propose a farthest cosine sampling based on cosine distance. It can sample the most distinctive features and capture the long-range dependencies. Extensive experiments and analysis show that our method achieves state-of-the-art performance at four sites with a total of 2687 WSIs. Furthermore, we demonstrate that our model can generalize to two unseen sites with 229 WSIs. Code is available at: https://github.com/boyden/PointTransformerFL

Introduction

Hematoxylin and eosin (HE)-stained whole slide images (WSIs) are now being used beyond visible tasks by applying deep learning methods (Lu et al. 2021; Shao et al. 2021; Li et al. 2022). These images contain subtle molecular characteristics that can be inferred using deep learning (Kather et al. 2020; Farahmand et al. 2022; Lu et al. 2022c). In breast cancer, accurately predicting human epidermal growth factor receptor 2 (HER2) status is crucial for guiding anti-HER2 treatment decisions (Oh and Bang 2019). Routinely, pathologists rely on HE-stained WSIs for breast cancer diagnosis, followed by specialized immunohistochemistry (IHC) and/or costly in-situ hybridization (ISH) techniques (Wolff et al. 2018) to determine HER2 status. By utilizing deep learning, we can predict HER2 status from broadly accessible HE-stained WSIs without requiring IHC and/or ISH.

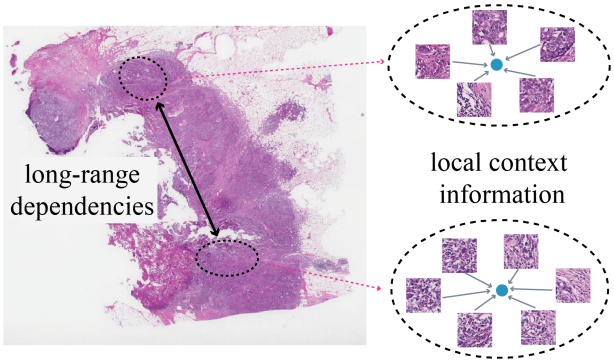

Figure 1: Local context information and long-range dependencies are both essential for WSI analysis

Achieving better WSI-level prediction requires large amounts of WSIs. Federated learning (FL) (McMahan et al. 2017) has already exhibited promising progress in WSI analysis (Lu et al. 2022b; Jiang, Wang, and Dou 2022; Ogier du Terrail et al. 2023). It can incorporate a large amount of multi-site WSIs without actual transportation of gigabyte-size WSIs and reduce the risk of data leakage. However, real-world WSIs exist non-independent and identically distributed (non-i.i.d.) scenarios. For HER2 classification, labeling imbalance and varying histological specimen preparation at different sites can adversely affect the overall performance. Although many studies have addressed the non-i.i.d. challenges (Hu et al. 2022; Guan and Liu 2023; Zhuang and Lyu 2023) in natural scenes, these methods remain a major gap in real-world WSIs compared to centralized learning.

In the site-end feature representation, WSIs are cut into patches, and these patches' local context information and

*These authors contributed equally.
†Corresponding author.

long-range dependencies (as shown in Figure 1) are essential for WSI-level prediction, such as HER2 prediction (Kather et al. 2020; Lu et al. 2022c) and survival analysis (Chen et al. 2021b; Shen et al. 2022; Shao et al. 2023). For HER2 prediction, HER2-positive patches may cluster in many separate regions of WSIs. Existing deep learning methods either treat these patches as instances using multi-instance learning (MIL) methods (Lu et al. 2021; Shao et al. 2021, 2023), or structure the patches into a graph using graph neural networks (GNNs) (Chen et al. 2021a; Lu et al. 2022c; Hou et al. 2022). However, the MIL-based methods lack the ability to model the local contextual information while the graph-based methods may struggle to capture long-distance dependence (Xu et al. 2018) and need extra edge representation. Alternatively, the point neural network (Qi et al. 2017a,b) can treat each patch as a point with inherent position information in the Euclidean space, and hence effectively model the local context by considering the position information. Moreover, it is permutation invariant and has demonstrated proficiency in aggregating features and representing long-range dependencies (Guo et al. 2021; Lu et al. 2022a), making it well-suited for WSI analysis.

In this paper, we introduce a PointTransformerDDA+ to represent both the local context and the long-range dependencies. Through the point transformer block, it can capture and aggregate the local information by employing attention mechanisms enriched with position information. We also propose a novel Farthest Cosine Sampling (FCS) to capture the long-range dependencies and gather the most distinctive features based on their cosine distance. To mitigate federated learning's label-imbalance of multi-site, we present a dynamic distribution adjustment (DDA) method for a well-initialized model. It includes a distribution adjustment strategy and an auxiliary classifier. The DDA allows the resampling of labels to the same imbalance ratio initially and then dynamically adjust to the real imbalance ratio for each site without degrading the feature representation.

Our main contributions can be summarized as follows:

- Unlike MIL models or graph models, we pioneer the use of point transformer for WSI analysis, which effectively captures both local context and long-range dependencies.
- Our proposed FCS can capture the long-range dependencies, leading to the most distinct feature aggregation.
- The proposed DDA mitigates the multi-site class imbalance issue, thereby enhancing model generalization.
- Extensive experiments on the largest WSI dataset to date for HER2 prediction in breast cancer demonstrate that our method achieves state-of-the-art performance in four sites (2687 WSIs) and two unseen sites (229 WSIs).

Related Work

In this section, we briefly review relevant works on WSI classification and federated learning in WSI analysis.

Whole Slide Image Classification

Recent works have used either MIL-based or graph-based methods for WSI classification, including molecular biomarkers such as HER2 status prediction (Farahmand et al. 2022; Lu et al. 2022c). The MIL-based methods commonly leverage attention mechanisms (Ilse, Tomczak, and Welling 2018; Chen et al. 2020; Lu et al. 2021) or transformers (Shao et al. 2021; Shen et al. 2022) to capture the long-distance dependence among instances. Regarding the local spatial relationship, DSMIL (Li, Li, and Eliceiri 2020) simply extracts feature from different scales and concatenate them among scales, which do not consider the local information in a specific scale. TransMIL (Shao et al. 2021) models the spatial relationship among patches via transformers with conditional position encoding; however, the positions used are not based on the actual Euclidean space. Graph-based models (Hamilton, Ying, and Leskovec 2017; Xu et al. 2019; Lee et al. 2022) are intrinsically designed to capture local information by a graph structure. In WSI analysis, Patch-GCN (Chen et al. 2021a) regards patches as 2D point clouds while still employing GNN to analyze WSIs. SlideGraph+ (Lu et al. 2022c) also constructs a graph based on position information and uses edge convolution (Wang et al. 2019) to model local neighbor features, leading to a SOTA HER2 prediction performance. However, the long-term dependency may limit the further improvement of GNNs in WSI analysis. While the permutation-invariant point neural network (Qi et al. 2017b; Zhao et al. 2021; Lu et al. 2022a) can capture both the local context and long-range dependencies, few studies have focused on WSI classification.

Federated Learning in WSI Analysis

Federated learning (McMahan et al. 2017; Guan and Liu 2023) can facilitate the training of data-driven models using multi-site WSIs. HistFL (Lu et al. 2022b) collaborates multi-sites WSI with attention MIL model and differential privacy for cancer subtype and survival prediction. Also, TNBC-FL (Ogier du Terrail et al. 2023) employs federated learning for predicting treatment outcomes in the rare subtype of breast cancer. However, general non-i.i.d. issues like skewed label distribution impedes the performance of federated learning. In natural scenes, several works replace batch normalization with group normalization (Hsieh et al. 2020), layer normalization (Du et al. 2022) or even remove the normalization layer (Zhuang and Lyu 2023) to address the problems by reducing the external covariate shift. FedProx (Li et al. 2020) introduces a regularization function to guarantee robust convergence of model in non-i.i.d. data. Additionally, FedMGDA (Hu et al. 2022) regards multi-site federated learning as a multi-objective optimization problem, aiming to converge to Pareto stationary solutions. With gradient normalization named FedMGDA+ (Hu et al. 2022), it can increase the model's robustness. Despite the progress made in federated learning, there still remains a gap between federated learning and centralized learning in real WSIs and further improvements are still needed.

Methodology

In this section, we start by introducing the problem definition of HER2 status prediction using a point transformer with federated learning. Then we describe the main components of the framework, including point feature extraction,

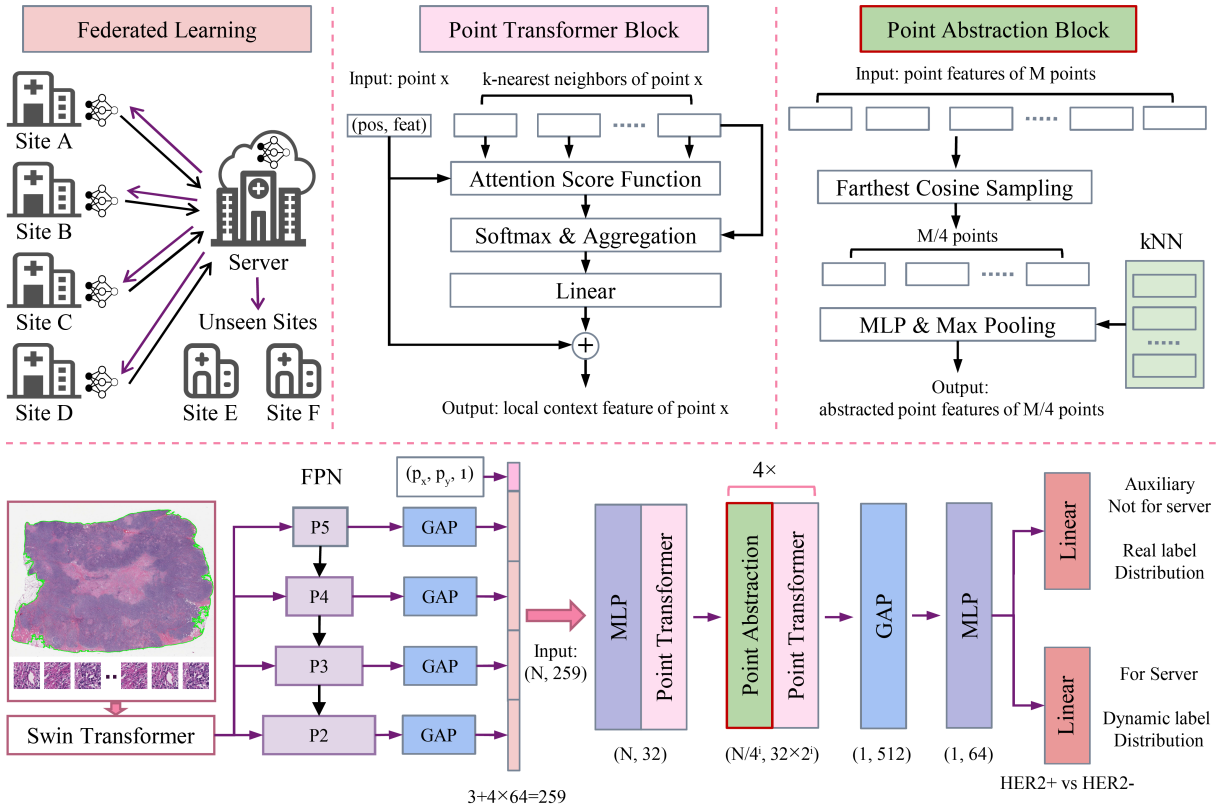

Figure 2: Overview of the point transformer for predicting HER2 status from whole slide images in a federated learning framework. $4\times$ represents that the corresponding blocks are repeated 4 times. In the i_{th} block, the output shape of point features is $(N/4^i, 32 \times 2^i)$ and N represents the total point numbers and is set to 1024. FPN: feature pyramid network, GAP: global average pooling, MLP: multilayer perceptron.

point transformer block, point abstraction block, and federated learning with dynamic distribution adjustment. Figure 2 illustrates the overall pipeline of our proposed framework.

Preliminaries

Suppose that we have M sites for HER2 status prediction, our goal is to accurately determine whether a given Whole Slide Image (WSI) is HER2-positive (HER2+) or HER2-negative (HER2-). For the i_{th} site, it contains a labeled point dataset $\mathcal{P}_i = \{(\mathcal{X}_n, y_n) \mid n \in (1, ..., |\mathcal{P}_i|)\}$, where $y_n \in \{0, 1\}$ is the corresponding HER2- and HER2+ status. Within this dataset, $\mathcal{X}_n = \{x_{n,1}, x_{n,2}, ..., x_{n,|\mathcal{X}_n|}\}$ represents a point set in the n_{th} whole slide images, where a point $x_{n,k} \in \mathbb{R}^{3+d}$ is a feature vector with 3-dim coordinates and d-dim point features. To determine HER2+ status from a point set $\mathcal{X}$ while considering data privacy, we employ federated learning to minimize the global cost over all sites:

$$\arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = \sum_{i=1}^{M} \frac{|\mathcal{P}_i|}{|\mathcal{P}|} \mathcal{L}_i(\mathcal{P}_i; \mathbf{W}), \quad (1)$$

where $|\mathcal{P}| = \sum_{i=1}^{M} |\mathcal{P}_i|$ is the total number of WSIs across all sites. For the i_{th} site, the cost can be calculated by:

$$\mathcal{L}_i(\mathcal{P}_i; \mathbf{W}) = \frac{1}{|\mathcal{P}_i|} \sum_{(\mathcal{X}_k, y_k) \in \mathcal{P}_i} \ell(f_c(f_h(\mathcal{X}_k)), y_k; \mathbf{W}), \quad (2)$$

where ℓ is the loss function, $f_h : \mathcal{X} \mapsto \mathbb{R}^d$ is a point set embedding function, and $f_c : \mathcal{R}^d \mapsto \mathbb{R}$ is the final classifier function. WSIs from real sites have a general non-independent and identically distributed (non-i.i.d.) scenario where each site has different HER2 status distributions. For the i_{th} site, we denote the class imbalance ratio as $\gamma_i = \frac{|\mathcal{P}_i^-|}{|\mathcal{P}_i^+|}$, where $|\mathcal{P}_i^-|$ and $|\mathcal{P}_i^+|$ represent to the number of HER2- and HER2+ WSIs within the i_{th} site.

Point Feature Extraction

To extract point features from a WSI, we follow the CLAM (Lu et al. 2021) to preprocess and patch the WSIs with details in Appendix A. Each patch is treated as a point. The corresponding coordinates for each patch are also traced and represented as a tuple$(p_x, p_y, 1)$, where 1 represents all WSIs having the same z-coordinate. Then we input the patches into a nuclei segmentation network that is pretrained using Swin-Transformer (Liu et al. 2021) with FPN (Lin

et al. 2017) with four levels, represented by P_2, P_3, P_4, P_5 in Figure 2. The channel of FPN is set to 64 and the outputs from all four layers of FPN are averaged and concatenated with the point coordinates as the patch-level feature $x_n \in \mathbb{R}^{3+d}$, where 3 represents the coordinates and $d = 256$ represents the point feature. Thus for a WSI with a label y, we can obtain a point set $\mathcal{X} = x_1, x_2, ..., x_{|\mathcal{X}|}$, where $|\mathcal{X}|$ is the number of patches in a WSI. 1024 patches or points are randomly selected with uniform distribution to reduce the memory usage and improve computational efficiency.

Point Transformer Block

In our approach, we adopt the original point transformer block (Zhao et al. 2021) to capture and aggregate the local context information of each point with an effective attention mechanism. For the i_{th} point with its corresponding point feature x_i, position p_i. We represent its *k*-nearest neighborhood points (k=16) as a subset $\mathcal{X}(i) \subset \mathcal{X}$. Then we compute the attention score between point x_i and point subset $\mathcal{X}(i)$ by the attention score function α:

$$\alpha(x_i, x_j) = W_q(W_i x_i - W_j x_j) + PE(p_i, p_j), \quad (3)$$

where $x_j \in \mathcal{X}(i)$ and PE is a relative position encoding function defined as:

$$PE(p_i, p_j) = MLP(p_i - p_j). \quad (4)$$

MLP in the formula represents two linear layers with a ReLU activation function.

Afterward, we aggregate the localized feature around of point x_i and obtain the aggregated feature z_i with a softmax function S. Then the output y_i is computed by applying a residual connection between x_i and z_i:

$$z_i = \sum_{x_j \in \mathcal{X}(i)} S(\alpha(x_i, x_j)) \cdot (W_v x_j + PE(p_i, p_j)), \quad (5)$$

$$y_i = x_i + W_z z_i. \quad (6)$$

Point Abstraction Block

To effectively reduce the cardinality of a point set and capture the long-range dependencies without missing important points, we propose a novel sampling strategy named farthest cosine sampling (FCS), as an alternative to the farthest point sampling (FPS) (Qi et al. 2017b).

In scenarios where a WSI exhibits a majority of negative patches with only a few positive patches clustered together in a specific region, FPS may miss these positive patches, as depicted in Figure 3. Consequently, it can lead to false negative predictions of HER2 status. Contrary to FPS, we perform sampling in the feature space, not in the position space. For a point set $\mathcal{X}_1 = \{x_1, x_2, ..., x_M\}$ with M points, we define the cosine distance as the distance metric between two points $x_i, x_j \in \mathcal{X}_1$:

$$Dist(x_i, x_j) = 1 - \frac{x_i \cdot x_j}{\max(||x_i||_2 \cdot ||x_j||_2, 1e-8)}. \quad (7)$$

Using this distance metric, we iteratively select the $M/4$ farthest points based on Algorithm 1. Consequently, the FCS can effectively cover the most requisite patches and capture the long-range dependencies in the feature space for better prediction of HER2 status.

Algorithm 1: Farthest cosine sampling

Input: M points with feature $\mathcal{X}_1 = \{x_1, x_2, ..., x_M\}$.
Output: Sampled M/4 points $\mathcal{X}_2$.

initialize an empty sampling point set $\mathcal{X}_2 = \{\}$;
$x_{s,1}$ = RandomChoiceOne($\mathcal{X}_1$);
$\mathcal{X}_1 \leftarrow \mathcal{X}_1 \setminus \{x_{s,1}\}$; $\mathcal{X}_2 \leftarrow \{x_{s,1}\}$;
while $|\mathcal{X}_2| < M/4$ **do**
  // Using cosine similarity for distance metric
  $Dist(i,j) = 1 - \frac{x_i \cdot x_{s,j}}{\max(||x_i||_2 \cdot ||x_{s,j}||_2, 1e-8)}$;
  $\mathcal{D} = \{Dist(i,j); x_i \in \mathcal{X}_1, x_{s,j} \in \mathcal{X}_2\}$;
  $x_s \leftarrow \arg\max_{x_i \in \mathcal{X}_1}(\mathcal{D})$;
  $\mathcal{X}_1 \leftarrow \mathcal{X}_1 \setminus \{x_s\}$; $\mathcal{X}_2 \leftarrow \mathcal{X}_2 \cup \{x_s\}$;
end while
return $\mathcal{X}_2$

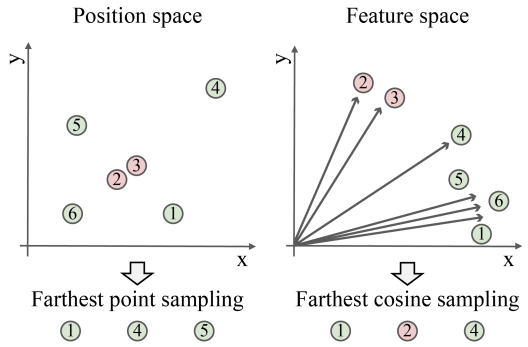

Figure 3: Difference between farthest point sampling and farthest cosine sampling. Light red: HER2+ points, light green: HER2- points.

After FCS, we obtain a sampled point subset $\mathcal{X}_2 = \{x_{s,1}, x_{s,2}, ..., x_{s,M/4}\}$. For each sampled point $x_i \in \mathcal{X}_2$, we define its *k*-nearest neighborhood points (k=16) on $\mathcal{X}_1$ as a point subset: $\mathcal{X}_2(i) \subset \mathcal{X}_1$. Subsequently, we group the feature from $\mathcal{X}_1$ onto $\mathcal{X}_2$ as y_i for each point $x_i \in \mathcal{X}_2$ using the following equation:

$$y_i = MaxPooling_{x_j \in \mathcal{X}_2(i)}(MLP(x_j)). \quad (8)$$

The MLP has two layers with each layer containing a linear transformation, batch normalization, and a ReLU activate function.

Point Classifier Block

After performing $4\times$ attention and abstraction operations, we obtain 4 abstract points with a grouped feature representation denoted as $F_g \in \mathbb{R}^{4\times512}$. By averaging the grouped feature, we derive the final WSI-level feature represented as $F_h \in \mathbb{R}^{64}$:

$$F_h = MLP(GAP(F_g)) = f_h(\mathcal{X}), \quad (9)$$

where MLP has two layers with each layer containing a linear transformation and a ReLU activation function. Following the MLP, a linear layer with a softmax function named

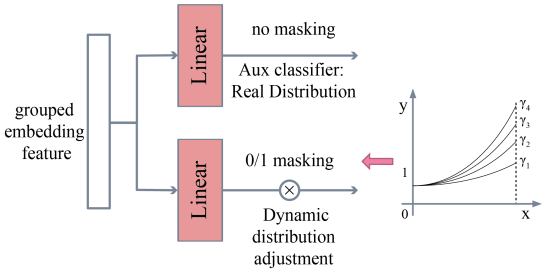

Figure 4: Dynamic distribution adjustment and auxiliary classifier for federated learning.

f_c outputs the final HER2 status probabilities p. The loss is calculated using cross entropy (CE) loss function, which is formulated as:

$$\ell(p, y) = -\frac{1}{|\mathcal{P}|}\sum_{i}^{|\mathcal{P}|}\sum_{c=0}^{1} y_c^i \log(p_c^i). \tag{10}$$

Federated Learning with Dynamic Distribution Adjustment

The WSIs from real sites exhibit a general non-independent and identically distributed (non-i.i.d.) scenario in which each site has a different label distribution, which can impact the performance of federated learning (Guan and Liu 2023). To mitigate this issue, we introduce a novel dynamic distribution adjustment strategy. It subsamples the majority of HER2- WSIs to the same label distribution in the initial training stage of federated learning. Then the balanced label distribution dynamically adjusts to the real distribution after progressive training. Specifically, we keep all the HER2+ WSIs and generated a 0/1 mask $M(k)$ for the HER2- WSIs. This mask is created using a Bernoulli distribution $\mathcal{B}$ with a probability of b_k at k_{th} epoch.

$$M_i(k) = \mathcal{B}(b_i^k), \tag{11}$$

$$b_i^k = \frac{1}{\gamma_i} + (1 - \frac{1}{\gamma_i}) \cdot \frac{e^{k/K} - 1}{e - 1} \in [\frac{1}{\gamma_i}, 1], \tag{12}$$

where γ_i is the i_{th} site's imbalanced ratio and K is the total optimization steps. The loss function can be formulated as:

$$\mathcal{L}_{cls} = M \cdot \ell(f_c(F_h), y) \tag{13}$$

$$= -\frac{1}{|\mathcal{P}_i|}[\sum_{i}^{|\mathcal{P}_i^-|} M_i(k) \log(p_0^i) + \sum_{i}^{|\mathcal{P}_i^+|} \log(p_1^i)], \tag{14}$$

where $|P_i^-|, |P_i^+|$ represent the number of HER2- and HER2+ WSIs.

In every epoch k, we note $b_i^k\gamma_i$ as the imbalance ratio involved in loss computation. Figure 4 shows that at epoch $k = 0$, all sites have $b_i^k\gamma_i = 1$ indicating that they share the same label distribution with an equal ratio between HER2- and HER2+. As the training progresses, the distribution gradually shifts towards each site's real distribution γ_i. By this strategy, the models are initially trained within a similar distribution, leading to a well-initialized model for accurately predicting HER2 status.

The above subsampling may arise a potential information loss problem, leading to insufficient feature representation. To address this concern, we add an auxiliary classifier that incorporates each site's real label distribution:

$$\mathcal{L}_{aux}(p, y) = -\frac{1}{|\mathcal{P}_i|}[\sum_{i}^{|\mathcal{P}_i^-|} \log(p_0^i) + \sum_{i}^{|\mathcal{P}_i^+|} \log(p_1^i)], \tag{15}$$

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{aux}. \tag{16}$$

This auxiliary classifier does not involved in the weights synchronization with the server model during the training of federated learning. Instead, it serves to guarantee the quality of feature representation as previous studies have demonstrated that models trained on imbalanced data can still learn high-quality feature representations (Kang et al. 2020; Lee, Shin, and Kim 2021). Detailed pseudo codes are shown in algorithm 2.

Experiments

Dataset and Experimental Settings

We evaluate our point transformer model for breast cancer HER2 status prediction using the largest WSI from six sites with a total of 2,916 WSIs. The sites are denoted as follows: Site A (TCGA-BRCA) (Network 2012), Sites B and C (internal hospitals with ethics committee approval), Site D (Conde-Sousa et al. 2022), and Sites E and F (Farahmand et al. 2022; Qaiser et al. 2017). Four sites participate in federated learning, while the remaining two sites serve as unseen data for external tests. Sites A, B, C, and D are split into training, validation, and test sets with a ratio of 6:1:3, as shown in Table 1. The splits are repeated five times, and the best model is selected based on the validation set in each split. The mean area under the ROC curve (AUC) is reported for the test set.

We refer to the point transformer with FCS as PointTransformer+, the variant with DDA as PointTransformerDDA, and the combined variant as PointTransformerDDA+.

WSIs	Federated Sites					Unseen Sites	
	Site A	Site B	Site C	Site D	Total	Site E	Site F
HER2−	669	672	332	306	1979	98	26
HER2+	118	214	172	204	708	93	12
Total	787	886	504	510	2687	191	38
γ	5.7	3.1	1.9	1.5	2.8	1.1	2.2
Train	472	532	302	306	1612	-	-
Val	79	88	50	51	268	-	-
Test	236	266	152	153	807	191	38

Table 1: WSIs and their HER2 status number in each site. $\gamma_i = \frac{|\mathcal{P}_i^-|}{|\mathcal{P}_i^+|}$ represents the imbalance ratio. WSIs are split into training (60%), validation (10%), and test (30%) set.

Experiments	Methods	Average	Site A	Site B	Site C	Site D
Ours	PointTransformerDDA+	**0.816** ± 0.019	**0.766** ± 0.025	**0.866** ± 0.021	**0.837** ± 0.036	**0.760** ± 0.046
	PointTransformerDDA	0.793 ± 0.013	0.730 ± 0.029	0.855 ± 0.013	0.804 ± 0.024	0.758 ± 0.022
	PointTransformer+	0.806 ± 0.015	0.752 ± 0.018	0.844 ± 0.008	0.823 ± 0.024	0.757 ± 0.043
Point-based	PointTransformer [1]	0.771 ± 0.012	0.717 ± 0.037	0.834 ± 0.026	0.776 ± 0.037	0.721 ± 0.035
	PointNet++ [2]	0.763 ± 0.017	0.696 ± 0.039	0.830 ± 0.033	0.746 ± 0.045	0.730 ± 0.042
MIL-based	CLAM-SB [3]	0.767 ± 0.032	0.712 ± 0.044	0.793 ± 0.037	0.766 ± 0.072	0.748 ± 0.022
	DSMIL [4]	0.693 ± 0.096	0.647 ± 0.065	0.738 ± 0.103	0.706 ± 0.080	0.675 ± 0.111
	TransMIL [5]	0.790 ± 0.019	0.739 ± 0.038	0.824 ± 0.021	0.805 ± 0.036	0.759 ± 0.040
	HistoFL [6]	0.757 ± 0.039	0.729 ± 0.048	0.776 ± 0.050	0.759 ± 0.076	0.733 ± 0.012
Graph-based	GraphSAGE [7]	0.711 ± 0.026	0.656 ± 0.053	0.735 ± 0.027	0.692 ± 0.044	0.685 ± 0.021
	Patch-GCN [8]	0.750 ± 0.037	0.700 ± 0.035	0.768 ± 0.062	0.766 ± 0.030	0.727 ± 0.047
	SlideGraph+ [9]	0.783 ± 0.019	0.736 ± 0.029	0.828 ± 0.012	0.804 ± 0.037	**0.785** ± 0.013

Table 2: Comparison of our model with other point, multi-instance, and graph-based models. [1] (Zhao et al. 2021), [2] (Qi et al. 2017b), [3] (Lu et al. 2021), [4] (Li, Li, and Eliceiri 2020), [5] (Shao et al. 2021), [6] (Lu et al. 2022b), [7] (Hamilton, Ying, and Leskovec 2017), [8] (Chen et al. 2021a), [9] (Lu et al. 2022c).

Implementation Details

Our models are implemented using PyTorch 1.12.0 on a workstation with an RTX 3090 GPU. The models are trained for 200 epochs with a learning rate of 1e-3 and L2 regularization of 1e-5. Point data augmentation is used with details in Appendix B. The learning rate warm-up is tuned for the first 10 epochs, followed by a cosine decay scheduler. Adam optimizer is adopted for weight updates.

Comparison with WSI Classification Methods

We include PointNet++ (Qi et al. 2017b), MIL-based models: CLAM-SB (Lu et al. 2021), DSMIL (Li, Li, and Eliceiri 2020), TransMIL (Shao et al. 2021), HistFL(Lu et al. 2022b), and graph-based models: GraphSAGE (Hamilton, Ying, and Leskovec 2017), Patch-GCN (Chen et al. 2021a), SlideGraph+ (Lu et al. 2022c) for comprehensive model comparison. All of the compared models are implemented with federated average settings.

Table 2 shows that point-based models offer a competitive performance compared to MIL-based and graph-based methods. The point-based models possess a unique advantage by effectively integrating both the local neighborhood features, similar to graph-based methods, and capturing the long-range dependencies, similar to MIL-based models. By introducing the novel FCS or/and DDA strategy, the point transformer achieves better performance compared to other models and PointTransformerDDA+ achieves the start-of-the-art AUC in the test set and three federated sites. Of note that TransMIL (Shao et al. 2021) also offers a high AUC compared to other related models. TransMIL also incorporates position encoding in the model, indicating that point position contributes to improved performance in predicting HER2 status. Further analysis of position information can be found in Appendix C.

Comparison with Federated Learning Methods

We also represent the point transformer's performance with different federated learning methods. Table 3 shows that our proposed PointTransformerDDA+ achieves the best total AUC among other methods and is the closest to the centralized training. Among the two proposed strategies FCS and DDA, FCS can capture the most discriminative features and DDA can mitigate the issue of non-i.i.d scenario; both of them can lead to a better federated learning performance. While GroupNorm (Hsieh et al. 2020) reaches higher performance at Site D, we observe that GroupNorm is sensitive to our data and relies on careful fine-grained group number selection with details in Appendix D. Moreover, although our models are not specifically designed for unseen scenarios, they still achieve commendable performance (AUC > 0.79) for two unseen sites, as shown in Table 4.

Federated Settings	Average	Site A	Site B	Site C	Site D
Centralization	0.823	0.722	0.839	0.824	0.819
PointTransformerDDA+	**0.816**	**0.766**	**0.866**	**0.837**	**0.760**
PointTransformerDDA	0.793	0.730	0.855	0.804	0.758
PointTransformer+	0.806	0.752	0.844	0.823	0.757
FedAVG [1]	0.771	0.717	0.834	0.776	0.721
FedGroupNorm [2]	0.783	0.733	0.832	0.775	**0.768**
FedProx [3]	0.788	0.759	0.836	0.800	0.719
FedMGDA [4]	0.773	0.723	0.818	0.773	0.741
FedMGDA+ [4]	0.780	0.734	0.813	0.804	0.738
FedWon [5]	0.774	0.716	0.824	0.777	0.728

Table 3: Comparison of different federated learning settings. [1] (McMahan et al. 2017), [2] (Hsieh et al. 2020), [3] (Li et al. 2020), [4] (Hu et al. 2022), [5] (Zhuang and Lyu 2023).

Unseen Sites	Site E	Site F
PointTransformerDDA+	0.793	0.791
PointTransformerDDA	0.795	0.802
PointTransformer+	**0.800**	**0.806**

Table 4: Performance of the point transformer in the unseen sites.

Ablation Studies

Farthest cosine sampling We first evaluate the effectiveness of FCS in the DDA and base federated average settings. We also assess FCS's impact at each site by training locally rather than employing a federated learning scheme. Table 5 shows that FCS can consistently improves performance in

Algorithm 2: Point transformer with federated learning

Input: M participating sites with point set $\mathcal{P}_m = (\mathcal{X}_m, y_m)$ and label imbalanced ratio γ_m where $m \in \{1, ..., M\}$. Optimization epochs: K, communication pace: E.
Output: Model's weights W_s.

initialize the point transformer's function for each site: f_h^m, f_c^m, f_{aux}^m;
initialize the same weights for the server and sites: $W_s^0, W_{s,1}^0, ..., W_{s,M}^0$;
for $k = 0$ **to** $K - 1$ **do**
  for $m = 1$ **to** M **do**
    $F_h^m = f_h^m(\mathcal{X}_m)$;
    // Dynamic distribution adjustment
    $M_m = \mathcal{B}(\frac{1}{\gamma_m} + (1 - \frac{1}{\gamma_m}) \cdot \frac{e^{k/K} - 1}{e - 1})$;
    $\mathcal{L}_m = M_m \cdot \ell(f_c^m(F_h^m), y_m) + \ell(f_{aux}^m(F_h^m), y_m)$;
    // Local site update
    $W_{s,m}^{k+1} \leftarrow W_{s,m}^k - lr \cdot \nabla_{W_{s,m}^k} \mathcal{L}_m$;
  end for
  if $(k + 1) \mod E = 0$ **then**
    // Server site update
    for each layer l in point transformer **do**
      if l is not the auxiliary classifier's layer **then**
        $W_s^{k+1,l} \leftarrow \sum_{m=1}^{M} \frac{|\mathcal{P}_i|}{|\mathcal{P}|} W_{s,m}^{k+1,l}$;
        $W_{s,m}^{k+1,l} \leftarrow W_s^{k+1,l}$;
      end if
    end for
  end if
end for
return W_s

Methods	FCS	Average	Site A	Site B	Site C	Site D
DDA	×	0.793	0.730	0.855	0.804	0.758
	✓	**0.816**	**0.766**	**0.866**	**0.837**	**0.760**
Base	×	0.771	0.717	0.834	0.776	0.721
	✓	**0.806**	**0.752**	**0.844**	**0.823**	**0.757**
NoFed	×	-	0.639	0.799	0.687	0.728
	✓	-	**0.659**	**0.831**	**0.711**	**0.729**

Table 5: Ablation experiments of farthest cosine sampling (FCS). NoFed: the model is trained at each site locally.

IHC Score 2+	Average	Site A	Site B	Site D
PointTransformerDDA+	0.712	0.688	0.733	0.723
PointTransformerDDA	0.703	0.668	0.710	0.735
PointTransformer+	0.747	0.726	0.748	0.730

Table 6: The AUC of our model in the IHC score 2+ subset. Site C is excluded due to no IHC 2+ WSIs .

all settings, including local training without federated learning. FCS can capture more discriminative features, leading to better performance. Of note, FCS only brings little benefit for HER2 status in Site D. This could be attributed to the fact that Site D primarily consists of biopsy WSIs with a lower number of total points compared to other sites. Consequently, the use of FPS sampling is sufficient to cover almost all the patches in Site D.

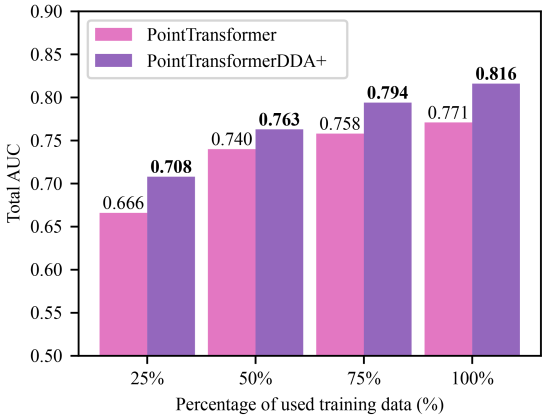

Figure 5: Performance compassion with different percentages of training WSIs.

Percentage of training WSIs Our model contains the largest number of WSIs to date for predicting HER2 status. However, real-world scenarios may lack sufficient WSIs. Therefore, we evaluate our model's performance by reducing the training WSIs to 75% (1209 WSIs), 50% (806 WSIs), 25% (403 WSIs). We exclude the use of 10% of the training WSIs as it results in less than 10 positive WSIs at each site, making it unsuitable for this experiment. Figure 5 shows that our model outperforms the base PointTransformer in all settings. The PointTransformerDDA+ with 50% of training WSIs achieve an AUC that is merely 0.008 lower than the PointTransformer model (0.763 vs 0.771) trained with 100% of the training WSIs. Moreover, with only 25% of training WSIs, the performance of PointTransformerDDA+ still outperforms DSMIL (Li, Li, and Eliceiri 2020) (0.708 vs 0.693).

IHC2+ Subset Analysis In real clinical scenarios IHC score 2+, pathologists cannot assess the HER2 status from IHC and require further expensive ISH tests. However, Table 6 shows that our model still achieves impressive performance with an average AUC > 0.7 in the test set. It can bring us an opportunity to reduce the reliance on ISH tests, thereby offering cost savings and faster HER2 status assessment.

Conclusion

Unlike MIL-based or graph-based methods, we regard a WSI as a point cloud with position information to derive the HER2 status from HE-stained WSIs, highlighting the effectiveness of point neural networks for WSI analysis. Specifically, a farthest cosine sampling is proposed to capture the long-range dependencies and aggregate most discriminative point features. Additionally, when utilizing federated learning, we proposed a dynamic distribution adjustment to mitigate the non-i.i.d. scenario of label imbalance in real-world WSIs. Extensive experiments have demonstrated the efficacy of our two components. Our models further achieve impressive performance in both unseen sites and IHC score 2+ subsets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62333022) and the Beijing Natural Science Foundation (No. JQ23034).

References

Chen, R. J.; Lu, M. Y.; Shaban, M.; Chen, C.; Chen, T. Y.; Williamson, D. F. K.; and Mahmood, F. 2021a. Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention*, 339–349. Springer.

Chen, R. J.; Lu, M. Y.; Wang, J.; Williamson, D. F.; Rodig, S. J.; Lindeman, N. I.; and Mahmood, F. 2020. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41: 757–770.

Chen, R. J.; Lu, M. Y.; Weng, W.-H.; Chen, T. Y.; Williamson, D. F.; Manz, T.; Shady, M.; and Mahmood, F. 2021b. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4025.

Conde-Sousa, E.; Vale, J.; Feng, M.; Xu, K.; Wang, Y.; Mea, V. D.; Barbera, D. L.; Montahaei, E.; Baghshah, M. S.; Turzynski, A.; Gildenblat, J.; Klaiman, E.; Hong, Y.; Aresta, G.; Araújo, T.; Aguiar, P.; Eloy, C.; and Polónia, A. 2022. HEROHE challenge: Predicting HER2 status in breast cancer from hematoxylin-eosin whole-slide imaging. *Journal of Imaging*, 8(8): 213.

Du, Z.; Sun, J.; Li, A.; Chen, P.-Y.; Zhang, J.; Li, H. H.; and Chen, Y. 2022. Rethinking normalization methods in federated learning. In *Proceedings of the 3rd International Workshop on Distributed Machine Learning*, 16–22.

Farahmand, S.; Fernandez, A. I.; Ahmed, F. S.; Rimm, D. L.; Chuang, J. H.; Reisenbichler, E.; and Zarringhalam, K. 2022. Deep learning trained on hematoxylin and eosin tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. *Modern Pathology*, 35(1): 44–51.

Guan, H.; and Liu, M. 2023. Federated learning for medical image analysis: A survey. arXiv:2306.05980.

Guo, M.; Cai, J.; Liu, Z.; Mu, T.; Martin, R. R.; and Hu, S. 2021. PCT: Point cloud transformer. *Computational Visual Media*, 7: 187–199.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 1025–1035.

Hou, W.; Yu, L.; Lin, C.; Huang, H.; Yu, R.; Qin, J.; and Wang, L. 2022. Hˆ2-MIL: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 933–941.

Hsieh, K.; Phanishayee, A.; Mutlu, O.; and Gibbons, P. 2020. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, 4387–4398. PMLR.

Hu, Z.; Shaloudegi, K.; Zhang, G.; and Yu, Y. 2022. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4): 2039–2051.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2127–2136. PMLR.

Jiang, M.; Wang, Z.; and Dou, Q. 2022. HarmoFL: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1087–1095.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 1–16.

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. 2020. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8): 789–799.

Lee, H.; Shin, S.; and Kim, H. 2021. ABC: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. In *Neural Information Processing Systems*, 7082–7094.

Lee, Y.; Park, J. H.; Oh, S.; Shin, K.; Sun, J.; Jung, M.; Lee, C.; Kim, H.; Chung, J.-H.; Moon, K. C.; et al. 2022. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, 1–15.

Li, B.; Li, F.; Liu, Z.; Xu, F.; Ye, G.; Li, W.; Zhang, Y.; Zhu, T.; Shao, L.; Chen, C.; et al. 2022. Deep learning with biopsy whole slide images for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study. *The Breast*, 66: 183–190.

Li, B.; Li, Y.; and Eliceiri, K. W. 2020. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14313–14323.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, 429–450.

Lin, T. Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 936–944.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9992–10002.

Lu, D.; Xie, Q.; Wei, M.; Gao, K.; Xu, L.; and Li, J. 2022a. Transformers in 3D point clouds: A survey. arXiv:2205.07417.

Lu, M. Y.; Chen, R. J.; Kong, D.; Lipkova, J.; Singh, R.; Williamson, D. F.; Chen, T. Y.; and Mahmood, F. 2022b. Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76: 102298.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570.

Lu, W.; Toss, M.; Dawood, M.; Rakha, E.; Rajpoot, N.; and Minhas, F. 2022c. SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer. *Medical Image Analysis*, 80: 102486.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 1273–1282. PMLR.

Network, T. C. G. A. 2012. Comprehensive molecular portraits of human breast tumors. *Nature*, 490: 61–70.

Ogier du Terrail, J.; Leopold, A.; Joly, C.; Béguier, C.; Andreux, M.; Maussion, C.; Schmauch, B.; Tramel, E. W.; Bendjebbar, E.; Zaslavskiy, M.; et al. 2023. Federated learning for predicting histological response to neoadjuvant chemotherapy in triple-negative breast cancer. *Nature Medicine*, 29: 135–146.

Oh, D. Y.; and Bang, Y. 2019. HER2-targeted therapies — a role beyond breast cancer. *Nature Reviews Clinical Oncology*, 17: 33–48.

Qaiser, T.; Mukherjee, A.; Reddy Pb, C.; Munugoti, S. D.; Tallam, V.; Pitkäaho, T.; Lehtimäki, T.; Naughton, T.; Berseth, M.; Pedraza, A.; et al. 2017. HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology*, 72: 227–238.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, 5105–5114.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; and Zhang, Y. 2021. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, 2136–2147.

Shao, Z.; Chen, Y.; Bian, H.; Zhang, J.; Liu, G.; and Zhang, Y. 2023. HVTSurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2209–2217.

Shen, Y.; Liu, L.; Tang, Z.; Chen, Z.; Ma, G.; Dong, J.; Zhang, X.; Yang, L.; and Zheng, Q. 2022. Explainable survival analysis with convolution-involved vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2207–2215.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Wolff, A. C.; Hammond, M. E. H.; Allison, K. H.; Harvey, B. E.; Mangu, P. B.; Bartlett, J. M.; Bilous, M.; Ellis, I. O.; Fitzgibbons, P.; Hanna, W.; et al. 2018. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline focused update. *Archives of Pathology & Laboratory Medicine*, 142(11): 1364–1382.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How powerful are graph neural networks? In *International Conference on Learning Representations*, 1–17.

Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, 5453–5462. PMLR.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zhuang, W.; and Lyu, L. 2023. Is normalization indispensable for multi-domain federated learning? arXiv:2306.05879.