Breast Cancer Classification From Digital Pathology Images via Connectivity-Aware Graph Transformer

Kang Wang, Feiyang Zheng, Lan Cheng, *Graduate Student Member, IEEE*, Hong-Ning Dai, *Senior Member, IEEE*, Qi Dou, and Jing Qin, *Senior Member, IEEE*

***Abstract*— Automated classification of breast cancer subtypes from digital pathology images has been an extremely challenging task due to the complicated spatial patterns of cells in the tissue micro-environment. While newly proposed graph transformers are able to capture more long-range dependencies to enhance accuracy, they largely ignore the topological connectivity between graph nodes, which is nevertheless critical to extract more representative features to address this difficult task. In this paper, we propose a novel connectivity-aware graph transformer (CGT) for phenotyping the topology connectivity of the tissue graph constructed from digital pathology images for breast cancer classification. Our CGT seamlessly integrates connectivity embedding to node feature at every graph transformer layer by using local connectivity aggregation, in order to yield more comprehensive graph representations to distinguish different breast cancer subtypes. In light of the realistic intercellular communication mode, we then encode the spatial distance between two arbitrary nodes as connectivity bias in self-attention calculation, thereby allowing the CGT to distinctively harness the connectivity embedding based on the distance of two nodes. We extensively evaluate the proposed CGT on a large cohort of breast carcinoma digital pathology images stained by Haematoxylin & Eosin. Experimental results demonstrate the effectiveness of our CGT, which outperforms state-of-the-art methods by a large margin. Codes are released on https://github.com/wang-kang-6/CGT.**

Manuscript received 6 August 2023; revised 26 November 2023 and 14 March 2024; accepted 20 March 2024. Date of publication 25 March 2024; date of current version 1 August 2024. This work was supported in part by the General Research Fund of Hong Kong Research Grants Council under project 15218521 and in part by the Theme-based Research Scheme of Hong Kong Research Grants Council under Project T45-401/22-N. *(Corresponding author: Jing Qin.)*

Kang Wang and Jing Qin are with the Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China (e-mail: kangwang@polyu.edu.hk; harry.qin@polyu.edu.hk).

Feiyang Zheng is with the School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China (e-mail: feiyangzheng@hust.edu.cn).

Lan Cheng is with the Big Data Bio-Intelligence Laboratory, Big Data Institute, The Hong Kong University of Science and Technology, Hong Kong, China (e-mail: chenglan@ust.hk).

Hong-Ning Dai is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China (e-mail: hndai@ieee.org).

Qi Dou is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: qidou@cuhk.edu.hk).

Digital Object Identifier 10.1109/TMI.2024.3381239

***Index Terms*— Tissue connectivity, tissue topology phenotyping, graph transformer, cancer classification, entity graph.**

I. Introduction

BREAST cancer is the most commonly diagnosed female cancer and ranks as the first for cancer mortality in women [1], [2]. Recent statistics by the American Cancer Society showed that breast cancer survival varies significantly by stage at diagnosis. The 5-year survival rates of USA patients diagnosed during 2012-2018 were >99% for stage I, 93% for stage II, 75% for stage III, and 29% for stage IV [3]. Early screening enables a timely risk assessment and expedites an optimal treatment plan [4], which has proved to be of great significance in reducing morbidity [5]. In clinical practice, microscopic analysis by pathology imaging is regarded as the "gold standard" for the final determination of breast cancer [6]. Manually screening these pathology images is, however, laborious, time-consuming, and error-prone [7]. To this end, automated classification approaches are highly demanded in clinical practice. It remains a challenging task due to (i) the complicated spatial patterns of cells in tissue micro-environment among different cancer categories [8]; (ii) the existence of inter-class similarities and intra-class variations [9], as illustrated in the top and bottom row in Fig. 1, respectively; and (iii) the extremely high resolution of digital pathology images [10], which may lead to huge computational costs.

Recent studies have demonstrated that graph neural networks (GNNs) are promising tools in breast cancer digital pathology image classification [11], [12]. Different from convolutional neural networks (CNNs), GNN-based methods take graphs built from pathology images as the input for cancer diagnosis, instead of directly performing on images, which not only more faithfully reflects the topological characteristics of the biological entities (*e.g.,* cells, tissues) but also has the potential to greatly reduce the computational costs. Most existing GNN-based methods reconstruct the graphs by tiling the pathology images into multiple smaller patches and setting them as graph nodes [13], [14]. However, such a strategy is incapable of capturing the topological relationships among the cells and/or tissues, which are important for an

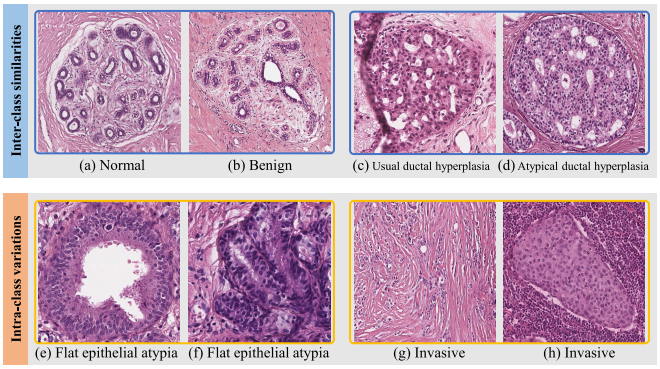

Fig. 1. Eight samples of digital pathology images for breast cancer classification, illustrating the complicated spatial patterns of cells in tissue micro-environment among different cancer categories. The upper row shows the similarities between the different categories. The bottom row shows the variations between the same categories.

intelligent model to apprehend their distributed patterns to yield satisfactory classification results. In addition, it usually suffers from selecting optimal patch resolution to obtain a balance between computational cost and effective feature extraction [10].

To address these limitations, a promising way is to construct entity graph [15], [16], [17] by segmenting biological entities in digital pathology images as graph nodes, such as cells and tissues. Hence, entity graphs can better reflect topological distributions and intrinsic interactions of these biological entities. In contrast to cell graphs, tissue graphs are able to more naturally capture relevant morphological regions, thereby being more scalable to large pathology images [18] and consequently reducing the computational cost significantly. Note that modern GNN-based methods for histological image analysis are dominated by message-passing GNNs. However, the message-passing mechanism can only propagate messages to neighborhood nodes. As a result, the dependencies among the distant nodes cannot be sufficiently explored.

In order to capture more global long-range dependencies among graph nodes in GNNs for further improving their representation capability, transformer-based GNNs, sometimes also termed as fully-connected GNNs or graph transformer, have been investigated and achieved promising performance [19], [20]. Although this fully-connected mechanism passes messages among all nodes, it largely neglects the input graph's actual connectivity, thereby losing, to some extent, the structural information of the graph. To tackle this shortcoming, some recently proposed Transformer-based GNNs attempt to integrate more structural information into graph representation to improve their capability [21], [22]. However, two limitations still remain. First, most existing Transformer-based GNNs usually incorporate structural information only by injecting a learnable positional encoding of each node into the input layer. Since such an injection is, nevertheless, disposable, this scheme still neglects the effect of the connected nodes in later training epochs, thus making structural information difficult to produce effective and profound influence on the generated representations. Second, when the transformer-based GNNs calculate the self-attention of a selected node, they equally treat the other nodes, ignoring the fact that their spatial distances to the selected node are essential for its representation.

To this end, we propose a novel connectivity-aware graph transformer (CGT) for breast cancer classification to sufficiently take connectivity attributes and spatial distance between tissue regions into account. The input of our method is the tissue graph constructed from pathology images, instead of directly using simple pixel information from pathology images. We summarize the main contributions of this work as follows:

- We propose the local connectivity aggregation method to add connectivity embedding to node features at every graph transformer layer, thereby making it capable of mapping comprehensive graph representations with the structural information of breast cancer subtypes.
- In light of the realistic intercellular communication mode, we propose to encode the spatial distance of a node-pair as the connectivity bias, to efficiently tame the connectivity embedding in the self-attention calculation.
- We evaluate the proposed CGT on two publicly annotated breast pathology image datasets. Extensive experiments demonstrate the effectiveness of our proposed method, which consistently outperforms state-of-the-art methods by a large margin.

II. Related Works

In this section, we briefly review the three related research directions including machine learning in pathology image classification, message-passing GNNs, and transformer-based GNNs.

A. Machine Learning in Classifying Pathology Images

Supervised learning has frequently been used in machine learning for histopathological image analysis [23], [24], [25]. However, current pathology images such as whole slide images (WSIs), contain gigapixels and often lack fine-grained annotations of tumor locations. As a result, deep neural networks [26], [27], [28] are prone to leverage a weakly supervised multiple instance learning (MIL) [29] for histopathological analysis, which processes pathology images in a patch-wise manner. Typically, the entire procedure of MIL can be broken down into two stages: 1) the first step is to create an instance-level classifier (*e.g.,* CNNs and ViTs) that maps split patches in square and fix-sized to a series of embedding vectors and calculates their positive probability; and 2) an aggregation network is then designed to create a bag-level feature vector and predict the final classification result of a pathology image. However, it is tricky to determine the optimal size of image patches for pathology images under different resolutions, thereby impeding these methods from capturing the topological relationships among the cells and/or tissues.

Currently, many researchers resorted to GNNs to describe tissue composition by incorporating morphology, topology, and interactions among biologically comprehensible entities, instead of square patches. For example, CGC-Net [30] was proposed to convert each large histology image into a graph, where cell morphology is embedded in the nodes. Anand et al. [31] proposed to use graph convolutional

networks (GCNs) for classifying patients into cancerous or non-cancerous groups on the Breast Cancer Histology Challenge (BACH) dataset [25]. Sureka et al. [32] modeled histology tissue as a graph of nuclei and employed robust spatial filtering (RSF) [33] with a GCN on the BACH dataset. Pati et al. [9] proposed a multi-level hierarchical entity graph representation of tissue specimens to model the hierarchical histological compositions. They also constructed a large cohort of Haematoxylin & Eosin stained breast pathology image datasets called BRACS. On the BRACS dataset, Jaume et al. [34] introduced a framework using entity-based graph analysis to provide pathologically understandable concepts, thereby easing pathologists' understanding of the graph decisions. Despite their favorable performance, the impacts of different GNN structures for pathology image classification have received insufficient attention to date.

B. Message-Passing GNNs

Message-passing graph neural networks (MP-GNNs) have made remarkable success for graph representation learning in diverse applications such as drug design [35], protein design [36], social network analysis [37], physics [38] and medical diagnosis [39]. After GCN [40] was first proposed to perform convolutions on the graph, Gilmer et al. [41] proposed a message-passing mechanism to enable the node to aggregate neighborhood information, which is the cornerstone of the recent emerging MP-GNNs [42], [43], [44]. After that, Pati et al. [45] proposed to use the Graph Isomorphism Network (GIN) [42], an instance of MP-GNNs, to subtype the breast cancer using tissue regions as nodes. Anklin et al. [18] proposed the SegGini, an MP-based GNN for weakly supervised segmentation, to segment tissue regions via the tissue graph. However, this vanilla message-passing mechanism results in the aggregation of only 1-hop node features by one GNN layer.

To improve the capability of graph representation, MP-GNNs either stacked a number of GNN layers or applied high-order GNN layers to progressively aggregate information from distant nodes. Nevertheless, MP-GNNs still have restricted expressiveness due to two major limitations (1) *over-smoothing* [46], [47], [48], [49], in which all node representations are prone to converge to a constant after passing through many stacked GNN layers; and (2) *over-squashing* [50], in which the distant node pairs cannot effectively interact using the message-passing mechanism in a graph since their interaction messages are directly compressed into the node features of fixed length. Hence, an urgent demand exists to exploit a new message-passing mechanism beyond neighborhood aggregation to yield more comprehensive graph representations.

C. Transformer-Based GNNs

Recent studies have shown that Transformers-based GNNs offer the potential to address the aforementioned issues owing to the self-attention mechanism proposed in [51]. Specifically, the self-attention mechanism of the Transformer-based GNNs enables two arbitrary nodes to interact with information. On the one hand, the vanilla Transformer-based GNNs consider the graph nodes as a discrete token sequence without nodes connecting relationships, empowering the model's global reasoning capability. On the other hand, this self-attention mechanism disregards the intrinsic graph structure, failing to identify nodes with similar structure (e.g., node degree) and capture the graph topology. This effect accounts for the reason why their performance was inferior to that of MP-GNNs in several tasks [19].

Therefore, it is promising to encode structural information to Transformer-based GNNs without the adjustment of the Transformer architecture. Graph Transformer [20] provided an early example of how to generalize the Transformer architecture to graphs, using Laplacian eigenvectors as an absolute encoding and computing attention on the immediate neighborhood of each node, rather than on the full graph. SAN [21] also used the Laplacian eigenvectors for computing an absolute encoding, but computed attention on the full graph, while distinguishing between true and created edges. Graphormer [22] proposed to encode some carefully selected graph theoretic properties as positional embeddings and attention bias, such as centrality measures and shortest path distances. Notably, these GNNs are designed for molecular structure and social network analysis, which may not be applicable to breast cancer classification. To the best of our knowledge, our proposed method is the first transformer-based GNN for breast cancer classification.

III. Methods

A. Preliminary

The task of breast cancer classification is, in principle, a multi-class classification problem. Let $G(V, E, H)$ denote an attributed-yet-undirected entity graph (*i.e.,*the tissue graph in our implementation) constructed from a pathology image, where $V = \{v_1, v_2, \cdots, v_N\}$ is the set of nodes, E is the set of edges, and N is the number of nodes ($N = |V|$). A set of node features are denoted by $H = \{h_1, h_2, \cdots, h_N\}, h_i \in \mathbb{R}^F$, where F is the feature dimension. An edge between two connected nodes $u, v \in V$ is denoted by $e_{u,v}$. The graph topology is described by a symmetric adjacency matrix $A \in \mathbb{R}^{N\times N}$, where $A_{u,v} = 1$ if $e_{u,v} \in E$. The neighborhood of node $v \in V$ is denoted by $\mathcal{N}(v) = \{u \in V | A_{u,v} = 1\}$.

Given a GNN for multi-class classification with the function space of $\mathcal{F}$, training this network is to find a classifier function $f \in \mathcal{F}$, which can effectively map the input graph representations of breast cancer data to appropriate predictions or scores. The classifier's role is to assign a label or score to each input graph, indicating the likelihood of belonging to a specific breast cancer category. The supervised learning for breast cancer classification can be defined by the following empirical risk optimization problem as follows:

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{N_t}\sum_{i=1}^{N_t}\mathcal{L}(f(x_i), y_i), \tag{1}$$

where x_i, y_i represent the input graph and the label of the i-th training example, respectively. The total number of training examples is denoted by N_t and $\mathcal{L}$ is the loss function for each training example. Each term in the sum represents the loss for

a specific training example, and the objective is to minimize the average loss over the entire training set.

B. Tissue Graph Construction

We consider tissue regions in a pathology image as nodes and construct a tissue graph rather than a cell graph for the two following reasons. First, cancer cells in tumors are embedded in complex tissues, namely *tissue micro-environment*, which consists of immune cells, stromal cells, and blood vessels [18], [52]. The interactions of these tissue regions in the tissue micro-environment exert different selective pressures on the evolution of cancerous regions. In this regard, pathologists mainly rely on observing the tissue distribution of the tissue micro-environment and studying its characteristics to determine the type and the grade of cancer. Second, a tissue graph has much fewer nodes than a cell graph, thereby greatly alleviating the computational cost [53]. In view of the merits, the tissue graph is constructed to feed our proposed CGT, consequently encouraging the CGT to efficiently capture and understand a high-level tissue micro-environment.

For a tissue graph, its nodes V and edges E denote tissue regions and their relations, respectively. There are two steps to identify tissue regions from a pathology image x. First, we use an unsupervised segmentation method, called simple linear iterative clustering (SLIC) algorithm [54], to initially group pixels into N_{sp} non-overlapping superpixels. Second, by comparing the RGB values of superpixels, several similar non-overlapping superpixels are merged into one homogeneous tissue region that captures meaningful tissue information (*e.g.,* epithelium, stroma, lumen, necrosis). The centroids of the merged tissue regions are considered as the nodes $V = \{v_1, v_2, \cdots, v_N\}$ of the tissue graph.

The feature representation of the tissue regions is obtained by another two-step procedure. First, we employ a ResNet34 [55], pre-trained on ImageNet dataset [56], to extract CNN features of each tissue region. Specifically, given the i-th tissue region, we catch all superpixels belonging to this tissue region. We iteratively feed a patch in size $h \times w$. This patch is centered around each superpixel centroid, into the pre-trained ResNet34 to compute the CNN feature of each superpixel. Second, the node feature h_i corresponding to the i-th tissue region is obtained by averaging the CNN features of its constituting superpixels. As such, a set of $H = \{h_1, h_2, \cdots, h_N\}, h_i \in \mathbb{R}^F$ constitutes the node features of the tissue graph, where F is the feature dimension.

Since the neighboring tissue regions are anticipated to interact biologically the most, they are set to be connected in the tissue graph. An edge $e_{u,v}$ is added between two nodes u, v when their represented tissue regions are adjacent according to the region adjacency graph [57]. The topology of the tissue graph is denoted by a binary adjacency matrix $A \in \mathbb{R}^{N \times N}$, where $A_{u,v} = 1$ if two nodes u, v are connected. Finally, the tissue graph of a pathology image is constructed as $TG(V, E, H)$.

C. CGT for Histopathological Representation

To represent the histopathological structure of the tissue graph, we propose the CGT to subtly incorporate connectivity information of the graph into the model. We devise the CGT architecture based on the original implementation of the classic Transformer encoder [51], where the encoder consists of multiple graph transformer layers (GTLs). Each GTL includes a multi-head self-attention (MHA) module and a position-wise feed-forward network (FFN). For the FFN sub-layer, we set the dimension of the input, the output, and the inner-layer to be the same as that reported in [22]. The calculation process of the standard GTL can be expressed as follows:

$$h'^{(\ell)} = \mathrm{MHA}\left(\mathrm{LN}\left(h^{(\ell-1)}\right)\right) + h^{(\ell-1)}, \tag{2}$$

$$h^{(\ell)} = \mathrm{FFN}\left(\mathrm{LN}\left(h'^{(\ell)}\right)\right) + h'^{(\ell)}, \tag{3}$$

where LN is the layer normalization function [58], $h^{(0)}$ is the input of the first GTL, *i.e.,* the node features of a given tissue graph $TG(V, E, H)$, and $h^{(\ell)}, \ell \in [1, 2, \cdots, L]$ denotes the output of the ℓ-th GTL, where L is the number of GTLs.

1) Connectivity Embedding With Local Connectivity Aggregation: For each node i, the self-attention mechanism in the MHA module can be used to calculate the semantic correlation between the node i and other nodes, implicitly reflecting the relation between node pairs and the structural information of a graph. However, such an attention calculation treats any graph as a fully-connected graph regardless of whether the node pairs are actually connected in the graph structure, neglecting the connectivity relationship in the tissue graph.

Therefore, to achieve better connectivity representation from the prior tissue topology knowledge, we propose to add connectivity embedding (CE) to the node feature for phenotyping tissue topology. Particularly, we initial learnable connectivity embedding $e_i^{(0)} \in \mathbb{R}^F$ at the input layer, which is assigned to each node based on itself degree as follows:

$$e_i^{(0)} = g(\mathrm{Deg}(v_i)), \tag{4}$$

where $\mathrm{Deg}(\cdot)$ denotes the function to calculate the node degree and $g : \mathbb{R}^1 \to \mathbb{R}^F$ denotes the learnable embedding function. In practice, function g creates a learnable matrix $\mathscr{C} \in \mathbb{R}^{N_m \times F}$ to map different degrees to a corresponding embedding, where N_m is the number of these learnable embeddings. Since the node with a different number of connected nodes has different topological information, the proposed CE enables our CGT to distinguish the connection capability of each node before calculating the self-attention in CGT.

We then add each CE to its corresponding node feature in the input layer as follows:

$$h_i^{(0)} = f_h(h_i, e_i^{(0)}) = h_i + \lambda e_i^{(0)}, \tag{5}$$

where f_h is the function of adding connectivity embedding and λ is a penalizing parameter to balance feature scale between h_i and e_i.

Note that the standard Transformer-based GNNs consider that each node is connected with all other nodes. Given a node embedding $h_i^{(\ell)}$, the update equation for a conventional GTL is defined as:

$$h_i^{(\ell)} = \mathrm{GTL}^{(\ell)}\left(h_i^{(\ell-1)}, \left\{h_j^{(\ell-1)}\right\}_{j \in \mathcal{V}}\right), \quad \ell \in [1, 2, \cdots, L] \tag{6}$$

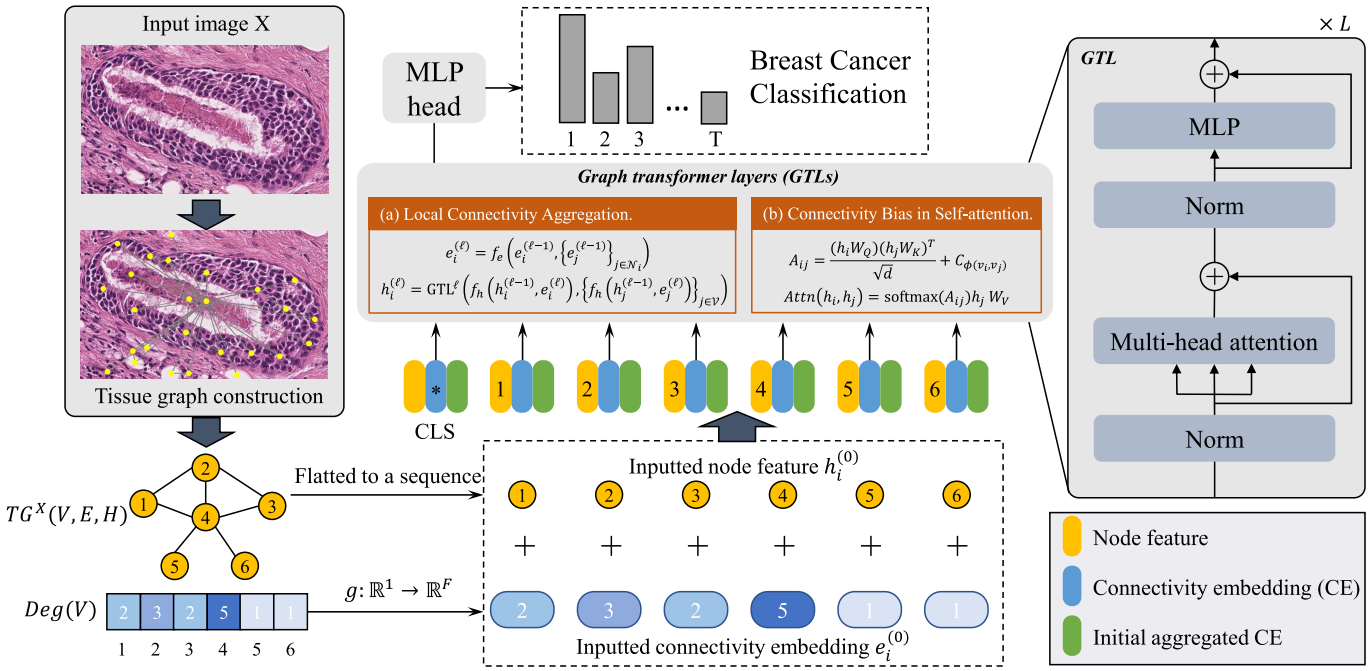

Fig. 2. The illustrative pipeline of how our CGT phenotypes the tissue topology connectivity for histopathological representation in breast cancer classification. The tissue graph $TG(V, E, H)$ is constructed from the input image as the input of the CGT. The nodes in the tissue graph are flatted to a sequence added with connectivity embedding, which is then fed to our proposed GTLs. To perform classification, we add an extra learnable classification token (CLS) to the sequence.

where $\mathrm{GTL}^{(\ell)}$ denotes the functions with learnable parameters of the ℓ-th GTL, and $\mathcal{V}$ denotes all nodes in V except node i.

Unlike the methods, which only add the structural information at the first GTL such as [20], [21], and [22], we propose the local connectivity aggregation (LCA) method to add aggregated CE to node feature at every GTL as follows:

$$e_i^{(\ell)} = f_e\left(e_i^{(\ell-1)}, \left\{e_j^{(\ell-1)}\right\}_{j\in\mathcal{N}_i}\right), \tag{7}$$

$$h_i^{(\ell)} = \mathrm{GTL}^{(\ell)}\left(f_h(h_i^{(\ell-1)}, e_i^{(\ell)}), \left\{f_h(h_j^{(\ell-1)}, e_j^{(\ell)})\right\}_{j\in\mathcal{V}}\right), \tag{8}$$

where $e_i^{(\ell)}, \ell \in [1, 2, \cdots, L]$ is the inputted CE of the ℓ-th GTL, f_e is the function to aggregate local connectivity, $\mathcal{N}_i$ is the neighborhood of the node i, and $h_i^{(\ell-1)}, h_i^{(\ell)}, e_i^{(\ell-1)}, e_i^{(\ell)} \in \mathbb{R}^F$.

2) Connectivity Bias in Self-Attention: Recent studies have shown that one important mode of intercellular communication is the release of soluble cyto- and chemokines in the cellular signaling processes [59], [60]. Once secreted, these signaling molecules diffuse through the surrounding medium and eventually bind to neighboring cell receptors whereby the signal is received, where the spreading speed depends on their spatial distance. However, most existing transformer-based GNNs neglect the spatial distance to calculate the self-attention of a selected node with the other nodes, which is not applicable to breast cancer classification. In light of this mode of intercellular communication, we propose to encode the spatial distance as the connectivity bias (CB) to self-attention calculation between two nodes, thereby allowing the CGT to distinctively harness the CE in the graph transformer architecture.

Algorithm 1 The Overall Process of CGT

Input: Tissue graph $TG(V, E, H)$ constructed from the image with breast cancer label y; Number of training epochs N_e;
Output: Predicted label $\hat{y}$

Parameter initialization;
for $e = 1, 2, \cdots, N_e$ **do**
  for all $v_i \in V$ **do**
    $e_i^{(0)} \leftarrow g(\mathrm{Deg}(v_i))$;
    $h_i^{(0)} \leftarrow f_h(h_i, e_i^{(0)})$;
    $\phi(v_i, v_j) \leftarrow \|p_i - p_j\|_2$;
  end for
  for $\ell = 1, 2, \cdots, L$ **do**
    for all $v_i \in V$ **do**
      $e_i^{(\ell)} \leftarrow f_e\left(e_i^{(\ell-1)}, \left\{e_j^{(\ell-1)}\right\}_{j\in\mathcal{N}_i}\right)$;
      $h_i^{(\ell)} \leftarrow \mathrm{GTL}^{(\ell)}(f_h(h_i^{(\ell-1)}, e_i^{(\ell)}), \left\{f_h(h_j^{(\ell-1)}, e_j^{(\ell)})\right\}_{j\in\mathcal{V}})$;
    end for
    Encode $C_{\phi(v_i,v_j)}$ as self-attention bias in the MHA module;
    Propagate MHA to the FFN;
    if $\ell == L$ **then**
      Export the CLS and obtain predicted label $\hat{y}$
    end if
  end for
  Update model parameters to minimize $\mathcal{L} = \mathcal{L}_{CE}(\hat{y}, y)$
end for

Let p_i, p_j denote the centroid positions of node i and node j, respectively. The spatial distance ϕ of node i and node j

is calculated as follows:

$$\phi\left(v_i, v_j\right) = \|p_i - p_j\|_2, \tag{9}$$

where $\|*\|_2$ is the function of computing 2-norm between two values.

Given the constructed tissue graph with node features $H = \{h_1, h_2, \cdots, h_N\}, h_i \in \mathbb{R}^F (i \in 1, 2, \cdots, N)$, we propose to calculate the self-attention CB-Attn between node i and node j in the MHA module as follows:

$$A_{ij} = \frac{\left(h_i W_Q\right)\left(h_j W_K\right)^T}{\sqrt{d}} + C_{\phi(v_i, v_j)}, \tag{10}$$

$$\text{CB-Attn}(H_{ij}) = \text{softmax}(A_{ij}) h_j W_V, \tag{11}$$

where A_{ij} denotes Query-Key product matrix and $C_{\phi(v_i,v_j)}$ is the function of mapping the node pair distance $\phi\left(v_i, v_j\right) \in \mathbb{R}^1$ to a learnable vector $v_d \in \mathbb{R}^d$, where d is the feature dimension of the learnable vector. Moreover, $W_Q \in \mathbb{R}^{F\times d}, W_K \in \mathbb{R}^{F\times d}, W_V \in \mathbb{R}^{F\times d}$ are three projection matrices corresponding to representation Query, Key, Value in the classic Transformer [51], respectively. Algorithm 1 shows the overall process of the proposed CGT.

IV. Experiments

A. Experimental Setup

1) Dataset:

a) BRACS dataset: We evaluated our CGT on a large cohort of breast cancer dataset termed BReAst Cancer Subtyping (BRACS) [9], which contains 4,391 breast tumor RoIs from 325 H&E breast carcinoma whole slide images. BRACS dataset was collected from 151 patients between 2019 and 2020, by board-certified pathologists of the Department of Pathology at the National Cancer Institute - IRCCS "Fondazione G. Pascale" in Naples (Italy). All slides were scanned with an Aperio AT2 scanner at 0.25μm pixel using a magnification factor of 40× resolution. The RoIs are annotated into seven categories by three pathologists as: Normal, Benign including Benign and Usual ductal hyperplasia (UDH), Atypical including Atypical ductal hyperplasia (ADH), and Flat epithelial atypia (FEA), Ductal carcinoma in situ (DCIS) and Invasive. The original split ratio of the training, validation, and test datasets at the RoI level is 3163:602:626.

b) BACH dataset: We also evaluated the proposed CGT on a publicly available digital pathology image dataset, *i.e.,* the Grand Challenge on BreAst Cancer Histology images BACH [25]. All images are acquired using a Leica DM 2000 LED microscope and a Leica ICC50 HD camera. These images are in RGB TIFF format and have a fixed size of 2048 × 1536 pixels, where the pixel scale is 0.42 × 0.42 μm. Two medical experts annotated the images into four categories, including Normal, Benign, DCIS, and Invasive. BACH contains the training and test images in the ratio of 400:100. All patients are from the Porto and Castelo Branco regions (Portugal). Cases are from Ipatimup Diagnostics and come from three different hospitals (Hospital CUF Porto, Centro Hospitalar do Tâmega e Sousa, and Centro Hospitalar Cova da Beira). The test data was collected from a completely different set of patients, ensuring a fairer evaluation of the methods.

In contrast, the training dataset and test dataset of BRACS are almost 10 and 6 times larger than those of BACH, respectively. BRACS has seven cancer subtypes that cover the whole histopathological spectrum of breast cancer. ADH and FEA categories in BRACS represent a significant diagnostic conundrum that frequently arises in clinical practice due to their high likelihood of developing into cancer. As shown in Table I, BRACS holds more variable typical and atypical hyperplasia subtypes, which are clinically more representative and resemble a realistic scenario of classifying breast cancer. Since transformer architecture is deemed to be data-eager over other deep learning networks, we conduct comparative experiments on BACH to further evaluate the classification performance of the proposed CGT on the lightweight dataset.

2) Implementation Details:

The implementation includes three processes: tissue graph construction, node distance computation, and breast cancer classification. We implement our methods using PyTorch [67] and the Deep Graph Library (DGL) [63]. In the process of tissue graph construction, the number of non-overlapping superpixels N_{sp} is chosen to be 500, and other parameters of the SLIC algorithm refer to the implementation in [9]. We use Cross Entropy Loss to train all methods with 10^{-3} learning rate. All methods were trained for 100 epochs using the Adam optimizer [68], 10^{-3} weight decay, and batch size as 4 by default. We let the number of GTLs be 6, the attention heads of each GTL be 6, and the hidden dimension of the node feature be 514, such that the total number of trainable parameters is in the range of 9M. Compared by the classification performance of four LCA functions (*i.e.,* sum, mean, max, min), we select the sum function to aggregate l-hop connectivity embedding in each GTL for the proposed CGT. Experiments are conducted on a single NVIDIA GeForce GTX 1080TI GPU with 11 GB memory. Table II reports the detailed model settings.

3) Competitors:

We compare our method against three groups of methods: message-passing GNNs, transformer-based GNNs, and vision transformers-based methods. Message-passing GNNs includes Patch-GNN [13], CG-PNA [43], TG-PNA [43], CGC-Net [30], and HACT-Net [9]. Transformer-based GNNs include GT [20], SAN [21], and Graphormer [22]. Vision transformers-based methods include HIPT [61] and CTransPath [62]. As mentioned before, Patch-GNN employs the patch graph as its input, whereas other GNNs use entity graphs. CG-PNA and CGC-Net employ the cell graph, and HACT-Net employs the hierarchical graph concatenated by the cell graph and tissue graph. Vision transformers-based methods directly use pathology images. The remaining competitors, as well as the proposed CGT, employ the tissue graph as their input. For a fair comparison, we adopted the same experimental settings when implementing these methods, and their hyperparameters were chosen to be the recommended values by their authors.

B. Experimental Results

1) Improvement Over SOTAs on BRACS Dataset Using Cross-Validation:

We first demonstrate our CGT works better than existing state-of-the-art (SOTA) methods for breast cancer classification. We provide classification accuracy to evaluate overall breast cancer classification performance on the BRACS

TABLE I

THE KEY DATA DESCRIPTION FOR THE BRACS DATASET FOR BREAST CANCER CLASSIFICATION, INCLUDING SUBTYPES DISTRIBUTION, STATISTICS OF TISSUE-GRAPH COMPONENTS, AND ROIS SPLIT STRATEGY. IN DETAIL, THERE ARE THREE DISTRIBUTION METRICS OF ROIS, INCLUDING THE NUMBER OF ROIS, THE NUMBER OF PIXELS (IN MILLIONS), AND THE RATIO OF MAX PIXEL AND MIN PIXEL. THERE ARE THREE METRICS TO DEPICT THE DISTRIBUTION OF TISSUE GRAPHS, INCLUDING THE NUMBER OF NODES AND EDGES, AND THE DEGREE OF NODES. MEAN AND STANDARD DEVIATION ARE ALSO PROVIDED

Class	RoIs			Tissue graphs			RoIs split		
	No. RoIs	No. Pixels	Max /Min pixel	No. Nodes	No. Edges	No. Degree	Train	Validation	Test
Normal	512	2.8±2.7	75.3	107±106	509±545	4.0±6.5	342	86	84
Benign	758	5.7±4.5	97.9	217±233	1012±1236	4.4±8.3	586	87	85
UDH	471	2.4±2.9	180.1	88±93	393±450	4.6±4.8	303	88	80
ADH	568	2.2±2.0	75.3	100±91	480±474	4.3±5.0	405	77	86
FEA	783	1.2±1.1	58.3	45±32	194±159	4.2±3.4	899	85	99
DCIS	749	5.0±5.0	128.6	225±217	1111±1123	4.5±7.2	562	97	90
Invasive	550	8.2±5.4	62.4	423±317	2025±1741	3.2±14.4	366	82	102
Total	4,391	3.9±4.3	235.6	172±217	815±1125	4.1±9.2	3,163	602	626

TABLE II

MODEL SETTINGS OF THE PROPOSED CGT

Model settings	CGT
GTL Layers	6
Node Hidden Dimension	514
FFN Inner-layer Dimension	256
Attention Heads	6
FFN Dropout	0.1
Attention Dropout	0.1
Embedding Dropout	0.1
Max Epochs	100
Peak Learning Rate	0.0002
Batch Size	4
Warm-up Steps	60K
Learning Rate Decay	Linear
Initial learning rate	0.001
Weight Decay	0.001

TABLE III

PERFORMANCE IMPROVEMENT COMPARED TO STATE-OF-THE-ART METHODS ON THE BRACS DATASET USING FIVE-FOLDER CROSS-VALIDATION. THE RESULTS INCLUDE MEAN VALUE AND STANDARD DEVIATION OF WEIGHTED F1 SCORES AND ACCURACY FOR THE TOTAL 7-CLASS CLASSIFICATION. RESULTS ARE PRESENTED IN %. THE BEST RESULT IS IN **BOLD**

Methods	Weighted F1		Accuracy	
	Mean	Std	Mean	Std
TG-PNA [43]	62.63	1.55	63.26	1.83
HACT-Net [9]	69.61	0.98	70.39	1.26
GT [20]	61.79	0.41	62.33	0.67
Graphormer [22]	70.89	1.27	71.36	1.55
HIPT [61]	70.29	0.98	71.01	1.26
CTransPath [62]	69.97	1.73	70.41	1.84
CGT (Ours)	**75.74**	**0.47**	**76.58**	**0.75**

dataset. Since the BRACS dataset has a large range of data variation in different classes (reported in Table I), we also utilize the weighted F1 score for performance evaluation. The weighted F1 score assigns different weights to each class based on their relative frequencies, so that the performance of minority classes is given more importance and is not overshadowed by the majority classes. The mean and standard deviation of per-class metrics are reported to indicate the distribution of classification results.

As shown in Table III, the classification result is produced by using five-fold cross-validation, *i.e.,* four-fold for training and one-fold for testing. The implementation details including the loss function, learning rate, the optimizer, the training epochs, and so on, were the same as ours (reported in Sec. IV-A.2). We can see from Table III that the proposed algorithm works better than SOTA methods including message-passing GNNs, transformer-based GNNs, and vision transformers-based methods according to weighted F1 score and accuracy. This experimental finding signifies that the proposed algorithm is able to boost breast cancer classification performance, even though the dataset holds numerous variable typical and atypical hyperplasia subtypes.

2) Improvement Over SOTA GNNs on BRACS Dataset: We compare the performance improvement of different methods on the BRACS dataset using the original data split, as shown in Table IV. We utilize the weighted F1 score on the BRACS dataset as the classification performance metric. Each model is trained three times by random weight initialization, exploiting potential classification performance sensitivity to initialization. The mean and standard deviation of per-class F1 scores are also provided.

From Table IV, we observe that our method consistently yields better performance than all competitors in weighted F1 score, implying that our method achieves the best breast cancer classification. Among message-passing GNNs, CG-PNA and TG-PNA perform better than Patch-GNN, indicating that traditional patch-wise GNNs are inferior to topological entity-based paradigms. TG-PNA obtains a higher performance gain than CG-PNA, demonstrating that using tissue graphs can obtain superior performance gains although their GNN backbones are identical such as the PNA herein. HACT-Net achieves the best results in message-passing GNNs but is still inferior to our CGT, confirming that the simple concatenation of cell graphs and tissue graphs is deficient to yield more comprehensive graph representations. Among transformer-based GNNs, GT, SAN, and Graphormer obtain a better performance than most of the message-passing GNNs except HACT-Net, thereby signifying the superiority of transformer-based GNNs. Our CGT surpasses the message-passing and

TABLE IV

PERFORMANCE IMPROVEMENT COMPARED TO DIFFERENT METHODS ON THE BRACS TEST DATASET, INCLUDING MEAN AND STANDARD DEVIATION OF PER-CLASS F1 SCORES AND WEIGHTED F1 SCORES FOR TOTAL 7-CLASS CLASSIFICATION. RESULTS ARE PRESENTED IN %. THE BEST RESULT IS IN **BOLD** AND THE SECOND BEST RESULT IS IN BLUE

Methods	Normal	Benign	UDH	ADH	FEA	DCIS	Invasive	Weighted F1
Patch-GNN [13]	52.53±3.27	47.57±2.25	23.67±4.65	30.66±1.79	60.73±5.35	58.76±1.15	81.63±2.17	52.10±0.61
CGC-Net [30]	30.83±5.33	31.63±4.66	17.33±3.38	24.50±5.24	58.97±3.56	49.36±3.41	75.30±3.20	43.63±0.51
CG-PNA [43]	58.77±6.82	40.87±3.05	46.82±1.95	39.99±3.56	63.75±10.48	53.81±3.89	81.06±3.33	55.94±1.01
TG-PNA [43]	63.59±4.88	47.73±2.87	39.41±4.70	28.51±4.29	72.15±1.35	54.57±2.23	82.21±3.99	56.62±1.35
HACT-Net [9]	61.56±2.15	47.49±2.94	43.60±1.86	40.42±2.55	74.22±1.41	66.44±2.57	88.40±0.19	61.53±0.87
GT [20]	59.48±1.19	46.43±2.46	23.12±0.04	35.93±2.60	71.51±0.31	63.78±2.24	85.57±3.88	56.64±0.18
SAN [21]	56.81±0.56	54.79±3.10	29.82±8.46	45.04±1.50	69.61±0.32	67.97±0.93	78.51±4.22	58.63±0.49
Graphormer [22]	58.51±4.25	**58.74±0.56**	37.11±7.80	47.87±1.82	68.49±1.11	69.40±1.96	83.93±2.85	61.70±1.11
CGT (Ours)	**65.56±0.42**	56.71±1.83	**47.12±0.29**	**50.47±1.15**	**75.06± 0.06**	**74.78±0.36**	**89.59±0.36**	**66.54±0.43**

TABLE V

PERFORMANCE IMPROVEMENT COMPARISON BETWEEN DOMAIN EXPERT PATHOLOGISTS AND PROPOSED CGT ON THE BRACS TEST DATASET, INCLUDING MEAN AND STANDARD DEVIATION OF PER-CLASS F1 SCORES, WEIGHTED F1 SCORES, AND WEIGHTED ACCURACY FOR TOTAL 7-CLASS CLASSIFICATION. RESULTS ARE PRESENTED IN %. THE BEST RESULT IS IN **BOLD**

Method	Normal	Benign	UDH	ADH	FEA	DCIS	Invasive	Weighted F1	Accuracy
Pathologist 1	67.53	53.92	41.90	36.00	19.13	71.59	94.00	55.30	56.71
Pathologist 2	47.83	52.94	25.00	35.37	65.22	68.00	94.00	57.07	57.99
Pathologist 3	39.66	49.59	49.43	42.29	54.12	65.19	89.47	56.71	56.55
Pathologist statistics	51.57±11.70	52.15 ±1.85	38.78 ±10.22	37.89 ±3.12	46.16 ±19.64	68.26 ±2.62	**92.49 ±2.14**	56.36 ±0.76	57.08 ±0.64
CGT (Ours)	**65.56±0.42**	**56.71±1.83**	**47.12±0.29**	**50.47±1.15**	**75.06± 0.06**	**74.78±0.36**	89.59±0.36	**66.54±0.43**	**67.32±0.35**

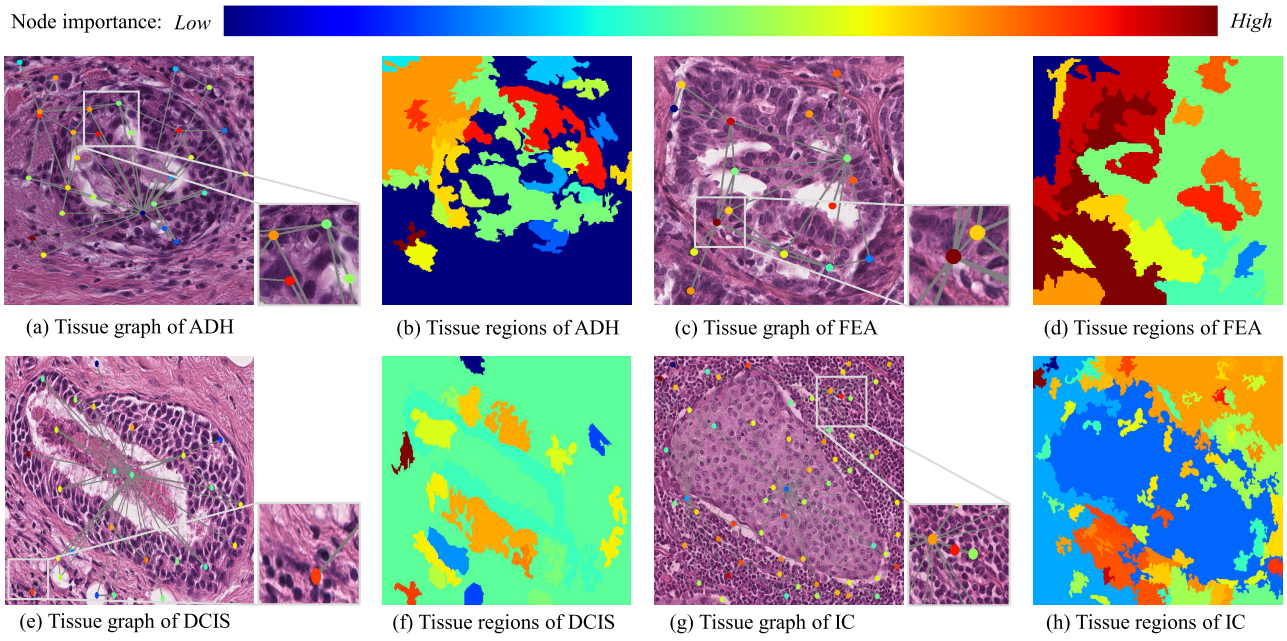

Fig. 3. The interpretation of our CGT using GraphGradCAM with four sample tissue graphs for 7-class breast cancer classification on the BRACS dataset. The sample tissue graphs with GraphGradCAM maps include (a) ADH, (c) FEA, (e) DCIS, and (g) IC, where partially enlarged figures are also provided. (b) (d) (f) (h) show the corresponding tissue regions with different importance scores.

transformer-based GNNs, demonstrating its better efficacy to phenotype the tissue topology for breast cancer classification.

As shown in Table V, the performance improvement of our proposed CGT is compared with three domain expert pathologists on the BRACS test dataset. We report the mean and standard deviation of per-class F1 scores, weighted F1 scores, and weighted accuracy for each independent pathologist. The united statistics of the three pathologists are also presented to benchmark our CGT. It is obvious that the classification performance of our CGT exceeds all domain expert pathologists in subtyping normal, ADH, and FEA. The CGT achieves comparable performance in classifying Benign, UDH, and DCIS categories. Compared with pathologists' statistics, the lower standard deviations of our CGT signify its superior and stable classification performance in breast cancer.

TABLE VI
PERFORMANCE IMPROVEMENT COMPARED TO THE ENSEMBLE AND SINGLE NETWORKS USING THE BACH TRAIN AND TEST DATASET. WE REPORT THE MEAN AND STANDARD DEVIATION OF ACCURACY FOR A TOTAL 4-CLASS CLASSIFICATION. RESULTS ARE PRESENTED IN %

Network types	Methods	Accuracy
Ensemble networks [25]	Wang et al. [63]	95.00
	Marami et al. [64]	94.00
	Chennamsetty et al. [65]	87.00
	Brancati et al. [66]	86.00
Single networks	HACT-Net [9]	91.00
	SAN [21]	89.00
	Graphormer [22]	90.00
	CGT (Ours)	92.00

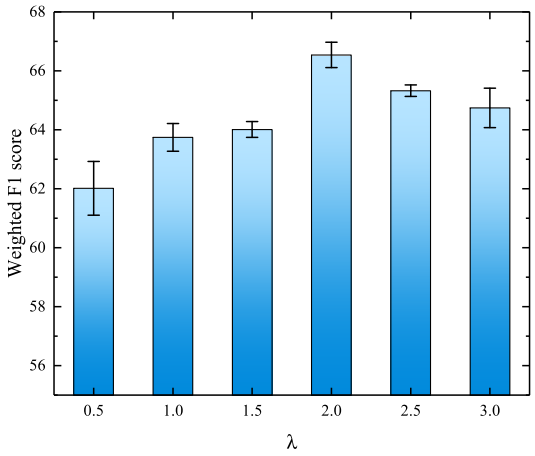

Fig. 4. The classification performance comparison of the proposed CGT variants using different values of hyperparameter λ on the BRACS test dataset. The histogram includes the mean and standard deviation of weighted F1 scores for the 7-class classification.

3) Improvement Over the Ensemble and Single Networks on BACH Dataset: We further compare the performance improvement of different methods on the BACH dataset, as shown in Table VI. The network types of these competitors are divided into two groups: ensemble networks and single networks [25]. Our prediction results are yielded by the organizers of the BACH challenge for a pair comparison. It can be seen that our CGT achieves the best classification performance among single networks and comparable accuracy in ensemble networks. The comparative result validates the breast cancer classification capability of the proposed CGT on the lightweight dataset. Although ensemble networks adopt multiple networks to generate prediction results, they also require more computation resources than single networks. The result demonstrates that our CGT can accurately classify breast cancer subtypes with a modest computation cost.

4) Explainability to Pathologists: To provide the explanations of our CGT to pathologists, a post-hoc gradient-based feature analysis approach termed GraphGradCAM [69] is adopted to highlight the nodes in the tissue graph, as well as their corresponding regions. As shown in Fig. 3, the nodes and corresponding regions in tissue graphs are represented in different colors, according to their importance levels in the breast cancer classification. We observe in Figs. 3(b)(d)(f)(h) that Our CGT focuses on the necrotizing tissue regions and tumorous epithelium in the tissue graph while ignoring the less important cell stroma. Interestingly, this observation indicates that our CGT can mimic the realistic pathological diagnosis for breast cancer, where the pathologists highly rely on the presence and morphological features of breast lesions to diagnose the corresponding cancer subtypes [7]. These explanations also enable pathologists to locate the relevant diagnostically tissue regions, consequently holding the potential to help pathologists assess their cancerization risks.

TABLE VII
ABLATION STUDY RESULTS ON THE BRACS DATASET USING LOCAL AGGREGATION AND ATTENTION BIAS. #PARAM IS THE NUMBER OF NETWORK PARAMETERS. RESULTS ARE PRESENTED IN %. THE BEST RESULT IS IN **BOLD**

Local aggregation			Attention bias		#param.	Weighted F1
CE	LCA	LFA	CB	SPD		
-	-	-	-	-	7904741	54.44±0.89
✓	-	-	-	-	8431077	59.76±0.37
-	-	✓	-	-	7904741	41.38±1.35
-	-	-	✓	-	7907813	61.43±0.47
-	-	-	-	✓	7910885	56.28±2.75
✓	-	-	✓	-	8434149	63.31±1.24
✓	✓	-	-	-	8431077	62.03±0.81
✓	✓	-	-	✓	8437221	63.09±0.31
✓	✓	-	✓	-	8434149	**66.54±0.43**

5) Hyperparameter Selection: The hyperparameter λ given in Eq. (5) is utilized to balance the scale between node features and their connectivity embeddings. Aiming to determine its best value, we conduct experiments with six λ values: 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 on the BRACS dataset. Fig. 4 shows that our method yields the highest mean of weighted F1 scores and its minimum weighted F1 score is still superior to that of other methods when $\lambda = 2$. Hence, we let λ be 2 in all other experiments.

C. Ablation Study

1) Local Connectivity Aggregation: To investigate the effect of the main components in our method: CE, LCA, and CB, we design several variants of the proposed CGT. The baseline method is the variant of the proposed CGT using no local aggregation or attention bias.

To compare with LCA, we design a variant using local feature aggregation (LFA), which only aggregates node features instead of CE. As reported in Table VII, adding CE to the node feature without LCA obtains a better result than the baseline. The variant of CE with LCA achieves a significant performance improvement over the variant of LFA (62.03±0.81 vs. 41.38±1.35), while the performance of the variant of LFA is inferior to that of the baseline (41.38±1.35 vs. 54.44±0.89). It indicates that only aggregating node features can offer limited benefits in classifying breast cancer, which is attributed to that an operation of local feature aggregation is equal to a message-passing GNN layer. The message-passing mechanism incurs the problem of over-smoothing and over-squashing based on

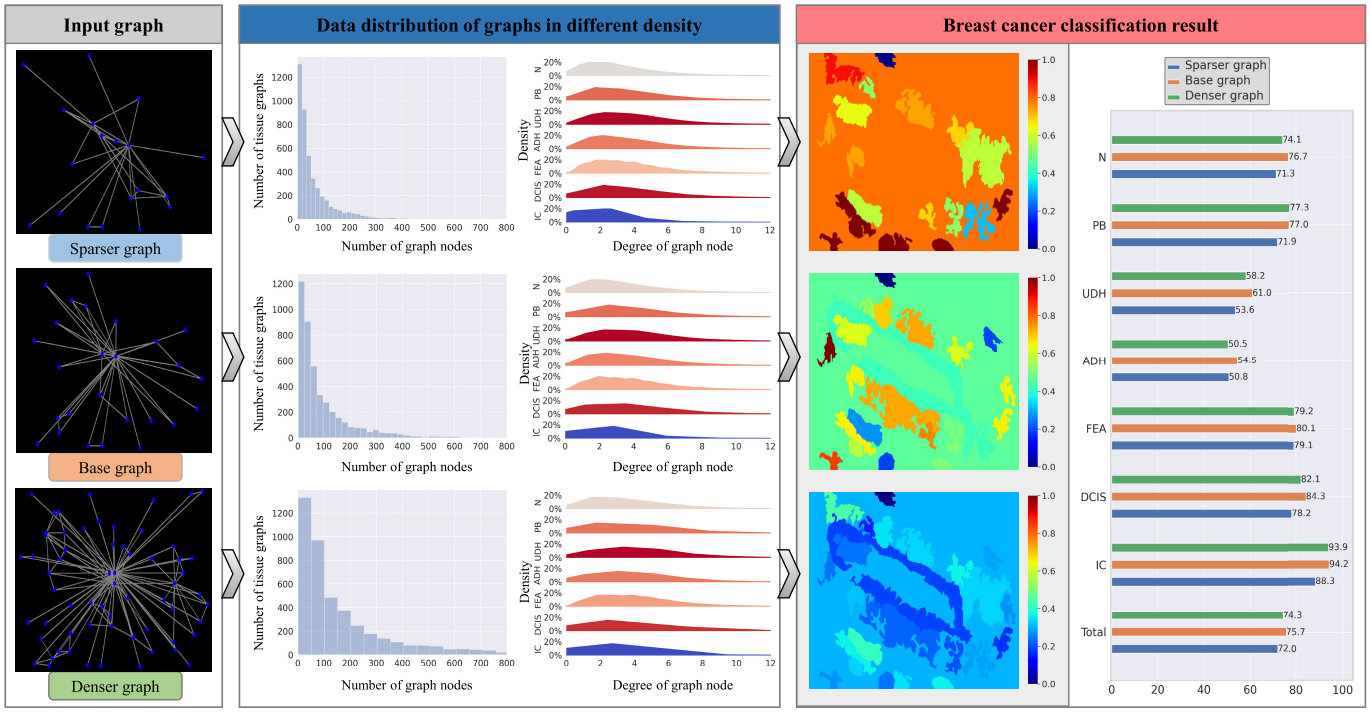

Fig. 5. The ablation result of the proposed CGT using the input graphs with diverse graph structures, *i.e.,* sparser, base, and denser graphs. The data distributions of these different types of graphs are provided in the middle region, including the number and degree of nodes. The region of the breast cancer classification result contains three corresponding interpretation maps and a mean F1 score histogram.

the aforementioned analysis, thereby leading to its inferior performance. Our full model (*i.e.,* CE+LCA+CB) yields a higher weighted F1 score than the variant using CE and CB, signifying that the proposed local connectivity aggregation is able to effectively map the structural information of tissue graphs to comprehensive histological representations. We can also see that the proposed LCA requires no extra network parameter, demonstrating the computation efficiency of the proposed CGT for breast cancer classification.

2) Connectivity Bias: To compare with CB, we design the variants by using shortest path distance (SPD) reported in [22] as another type of attention bias. Without any local aggregation operation, we can see from Table VII that the variant of CB achieves a better performance gain (61.43±0.47) than the baseline (54.44±0.89) and the variant of SPD (56.28±2.75). Also, the variant of CB outperforms most of the message-passing and transformer-based GNNs in Table IV, demonstrating that breast cancer classification indeed benefits from the proposed CB as the attention bias. It is observed that the variant of using CE and CB simultaneously works better than the variants of using independent connectivity attribute, *i.e.,* the variant of CE and the variant of CB. The result validates the effectiveness of adding adequate structural information of the tissue graph to transformer-based GNNs for capturing histopathological representation. Our full model (*i.e.,* CE+LCA+CB) surpasses other variants, conveying that the main components in our method are necessary and mutually reinforced for histopathological representation in breast cancer classification.

3) Graph Structure: Our proposed connectivity attributes including CE and CB, as demonstrated in the ablation study before, have significant effects on model performance for breast cancer classification. CE encodes graph topology by adding learnable connectivity embedding into node features, which is initialized by graph node degree. Additionally, CB aggregates spatial distance between two nodes by mapping the node pair distance to a learnable vector while calculating the self-attention among nodes. It is noted that our proposed CE and CB highly depend on the structure of the input graph, *e.g.,* the number and degree of nodes, and the denseness or sparseness of the graph.

Hence, we here conduct ablation experiments on the input graphs with different densities, to evaluate the effect of diverse graph structures on the proposed CGT. In detail, we construct three types of tissue graphs from the same pathology image on the BRACS dataset, by changing the number of non-overlapping superpixels N_{sp} to 300, 500, and 700 respectively, as described in Sec. III-B. Fig. 5 depicts three example input graphs, which are generated by the same DCIS image from Fig. 3(e). According to their node amounts, the constructed three types of tissue graphs are named sparser graph, base graph, and denser graph. The data distributions of the three types of tissue graphs are provided in Fig. 5, exhibiting their variance in the number and degree of nodes.

Except for the number of non-overlapping superpixels N_{sp}, the implementation details such as the loss function, learning rate, the optimizer, and the training epochs, were the same as ours as reported in Sec. IV-A.2. The breast cancer classification result is produced by using five-fold cross-validation, *i.e.,* four-fold for training and one-fold for testing. We can see from Fig. 5 that the base graph performs better than the sparser graph and denser graph, in terms of per-class F1 score and weighted F1 score. Although the sparser graph yields the lowest weighted F1 score, we observed

that it still outperforms the second-best model (72.0 vs. 70.89) in Table III. We provide three interpretation maps corresponding to the sparser graph, base graph, and denser graph, respectively, as shown in Fig. 5. It is observed that the importance scores of biological regions in the sparser graph are much closer and higher than those of other graphs. Meanwhile, the importance scores of biological regions in the denser graph are also close but lower. This experimental finding indicates that the sparser and denser graphs may contain less or more irrelevant information or noise, resulting in undesirable breast classification performance. The graph with the appropriate node amount could provide more representative features in pathological images, boosting the diagnosis performance of breast cancer. We thus let the number of non-overlapping superpixels N_{sp} be 500 in our proposed CGT.

V. Conclusion

Automated classification of breast cancer subtypes from digital pathology images is an extremely challenging task due to the complicated spatial patterns of cells in the tissue micro-environment. In this paper, we have presented CGT, a connectivity-aware graph transformer for breast cancer classification by phenotyping the topology connectivity of the tissue graph constructed from digital pathology images. The two contributions are summarized as: (i) Our CGT leverages the graph transformer architecture to add connectivity embedding at every graph transformer layer by using local connectivity aggregation, thereby mapping the comprehensive graph representations to breast cancer subtypes. (ii) The spatial distance is further encoded to the connectivity bias in self-attention calculation between two arbitrary nodes in a tissue graph, to efficiently capture and distinguish nodes' connectivity relationships.

We evaluated this novel network on the BRACS dataset which is a large cohort of annotated tissue RoIs from Haematoxylin & Eosin stained breast carcinoma digital pathology images. Our CGT is demonstrated to surpass state-of-the-art methods, indicating its better efficacy to phenotype the tissue topology for breast cancer classification. Further compared with pathologists' statistics, the lower standard deviations of our CGT signify its superior and stable classification performance. Various experiments are also conducted on a publicly available digital pathology image dataset BACH, signifying our CGT holds the potential to accurately classify pre-cancer subtypes with a modest computation cost. A comprehensive ablation study conveys that the main components (*i.e.,* connectivity embedding with local connectivity aggregation, and connectivity bias in our method are necessary and mutually reinforce histopathological representation in breast cancer classification.

By constructing entity graphs from different digital pathology images, our CGT can be potentially applied to the diagnosis of other cancer types. While the proposed CGT has shown promising results, there are still some limitations. WSI-level image classification would be ideal for clinical practice, but our proposed CGT is validated on patch-level image datasets. It is worth noting that our method can be easily modified as instance-level feature extractors for arbitrary MIL methods, highlighting the potential for WSI-level classification. In future research, we will consider addressing how clinical information, such as patient demographics, medical history, and gene sequence, can be integrated into our method, and how it can contribute to improving real-world diagnostic processes and patient outcomes.

References

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] A. G. Waks and E. P. Winer, "Breast cancer treatment: A review," *Jama*, vol. 321, no. 3, pp. 288–300, 2019.

[3] A. N. Giaquinto et al., "Breast cancer statistics, 2022," *CA, Cancer J. Clinicians*, vol. 72, no. 6, pp. 524–541, Nov. 2022.

[4] C. Allemani et al., "Global surveillance of cancer survival 1995–2009: Analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2)," *Lancet*, vol. 385, no. 9972, pp. 977–1010, Mar. 2015.

[5] R. A. Smith, V. Cokkinides, and H. J. Eyre, "American cancer society guidelines for the early detection of cancer, 2006," *CA, A Cancer J. Clinicians*, vol. 56, no. 1, pp. 11–25, Jan. 2006.

[6] N. Harbeck and M. Gnant, "Breast cancer," *Lancet*, vol. 389, no. 10074, pp. 1134–1150, 2017.

[7] J. G. Elmore et al., "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA*, vol. 313, no. 11, p. 1122, Mar. 2015.

[8] A. Myronenko, Z. Xu, D. Yang, H. R. Roth, and D. Xu, "Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 329–338.

[9] P. Pati et al., "Hierarchical graph representations in digital pathology," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102264.

[10] Z. Huang, H. Chai, R. Wang, H. Wang, Y. Yang, and H. Wu, "Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 561–570.

[11] B. E. Bejnordi et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017.

[12] A. Parvatikar et al., "Prototypical models for classifying high-risk atypical breast lesions," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 143–152.

[13] B. Aygüneş, S. Aksoy, R. G. Cinbiş, K. Kösemehmetoğlu, S. Önder, and A. Üner, "Graph convolutional networks for region of interest classification in breast histopathology," *Proc. SPIE*, vol. 11320, Aug. 2020, Art. no. 113200K.

[14] R. J. Chen et al., "Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 339–349.

[15] C. Gunduz, B. Yener, and S. H. Gultekin, "The cell graphs of cancer," *Bioinformatics*, vol. 20, no. 1, pp. I145–I151, Aug. 2004.

[16] W. Lu, S. Graham, M. Bilal, N. Rajpoot, and F. Minhas, "Capturing cellular topology in multi-gigapixel pathology images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 260–261.

[17] L. Studer, J. Wallau, H. Dawson, I. Zlobec, and A. Fischer, "Classification of intestinal gland cell-graphs using graph neural networks," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3636–3643.

[18] V. Anklin et al., "Learning whole-slide segmentation from inexact and incomplete labels using tissue graphs," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 636–646.

[19] V. P. Dwivedi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson, "Graph neural networks with learnable structural and positional representations," 2021, *arXiv:2110.07875*.

[20] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," 2020, *arXiv:2012.09699*.

[21] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, and P. Tossou, "Rethinking graph transformers with spectral attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21618–21629.

[22] C. Ying et al., "Do transformers really perform badly for graph representation?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28877–28888.
[23] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 34–42, Jan. 2018.
[24] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101813.
[25] G. Aresta et al., "BACH: Grand challenge on breast cancer histology images," *Med. Image Anal.*, vol. 56, pp. 122–139, Aug. 2019.
[26] K. Sirinukunwattana, N. K. Alham, C. Verrill, and J. Rittscher, "Improving whole slide segmentation through visual context—A systematic study," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, Granada, Spain. Cham, Switzerland: Springer, 2018, pp. 192–200.
[27] C. Mercan, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "From patch-level to ROI-level deep feature representations for breast histopathology classification," *Proc. SPIE*, vol. 10956, pp. 86–93, Mar. 2019.
[28] X. Wang et al., "TransPath: Transformer-based self-supervised learning for histopathological image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 186–195.
[29] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 10, 1997, pp. 570–576.
[30] Y. Zhou, S. Graham, N. A. Koohbanani, M. Shaban, P.-A. Heng, and N. Rajpoot, "CGC-Net: Cell graph convolutional network for grading of colorectal cancer histology images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 388–398.
[31] D. Anand, S. Gadiya, and A. Sethi, "Histographs: Graphs in histopathology," *Proc. SPIE*, vol. 11320, pp. 150–155, Sep. 2020.
[32] M. Sureka, A. Patil, D. Anand, and A. Sethi, "Visualization for histopathology images using graph convolutional neural networks," in *Proc. IEEE 20th Int. Conf. Bioinf. Bioengi. (BIBE)*, Oct. 2020, pp. 331–335.
[33] F. P. Such et al., "Robust spatial filtering with graph convolutional neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 884–896, Sep. 2017.
[34] G. Jaume et al., "Quantifying explainers of graph neural networks in computational pathology," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8102–8112.
[35] T. Gaudelet et al., "Utilizing graph machine learning within drug discovery and development," *Briefings Bioinf.*, vol. 22, no. 6, Nov. 2021, Art. no. bbab159.
[36] J. Ingraham, V. Garg, R. Barzilay, and T. Jaakkola, "Generative models for graph-based protein design," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 15820–15831.
[37] W. Fan et al., "Graph neural networks for social recommendation," in *Proc. World Wide Web Conf.*, May 2019, pp. 417–426.
[38] V. Bapst et al., "Unveiling the predictive power of static structure in glassy systems," *Nature Phys.*, vol. 16, no. 4, pp. 448–454, Apr. 2020.
[39] Y. Li, B. Qian, X. Zhang, and H. Liu, "Graph neural network-based diagnosis prediction," *Big Data*, vol. 8, no. 5, pp. 379–390, Oct. 2020.
[40] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2016.
[41] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1263–1272.
[42] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" 2018, *arXiv:1810.00826*.
[43] G. Corso, L. Cavalleri, D. Beaini, P. Liò, and P. Veličković, "Principal neighbourhood aggregation for graph nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 13260–13271.
[44] W. Hu et al., "Strategies for pre-training graph neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[45] P. Pati et al., "HACT-Net: A hierarchical cell-to-tissue graph neural network for histopathological image classification," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Cham, Switzerland: Springer, 2020, pp. 208–219.
[46] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 3538–3545.
[47] G. Li, M. Müller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9266–9275.
[48] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, "Measuring and relieving the over-smoothing problem for graph neural networks from the topological view," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 3438–3445.
[49] K. Oono and T. Suzuki, "Graph neural networks exponentially lose expressive power for node classification," 2019, *arXiv:1905.10947*.
[50] U. Alon and E. Yahav, "On the bottleneck of graph neural networks and its practical implications," 2020, *arXiv:2006.05205*.
[51] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[52] P. Martin-Gonzalez, M. Crispin-Ortuzar, and F. Markowetz, "Predictive modelling of highly multiplexed tumour tissue images by graph neural networks," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data*. Cham, Switzerland: Springer, 2021, pp. 98–107.
[53] Y. Zheng et al., "Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102308.
[54] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[57] F. K. Potjer, "Region adjacency graphs and connected morphological operators," in *Mathematical Morphology and Its Applications to Image and Signal Processing*. Boston, MA, USA: Springer, 1996, pp. 111–118.
[58] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[59] K. Francis and B. O. Palsson, "Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion," *Proc. Nat. Acad. Sci. USA*, vol. 94, no. 23, pp. 12258–12262, Nov. 1997.
[60] J. A. Papin, T. Hunter, B. O. Palsson, and S. Subramaniam, "Reconstruction of cellular signalling networks and analysis of their properties," *Nature Rev. Mol. Cell Biol.*, vol. 6, no. 2, pp. 99–111, Feb. 2005.
[61] X. Wang et al., "Transformer-based unsupervised contrastive learning for histopathological image classification," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102559.
[62] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16144–16155.
[63] M. Y. Wang, "Deep graph library: Towards efficient and scalable deep learning on graphs," in *Proc. ICLR Workshop Represent. Learn. Graphs Manifolds*, 2019.
[64] B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, and J. Zeineh, "Ensemble network for region identification in breast histopathology slides," in *Proc. Int. Conf. Image Anal. Recognit.* Cham, Switzerland: Springer, 2018, pp. 861–868.
[65] S. S. Chennamsetty, M. Safwan, and V. Alex, "Classification of breast cancer histology image using ensemble of pre-trained neural networks," in *Proc. Int. Conf. Image Anal. Recognit.* Cham, Switzerland: Springer, 2018, pp. 804–811.
[66] N. Brancati, M. Frucci, and D. Riccio, "Multi-classification of breast cancer histology images by using a fine-tuning strategy," in *Proc. Int. Conf. Image Anal. Recognit.* Cham, Switzerland: Springer, 2018, pp. 771–778.
[67] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.
[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[69] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10764–10773.