Medical Image Analysis 96 (2024) 103197

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Multi-scale relational graph convolutional network for multiple instance learning in histopathology images

Roozbeh Bazargani [a,*], Ladan Fazli [b,c], Martin Gleave [b,c], Larry Goldenberg [b,c], Ali Bashashati [d,e,1], Septimiu Salcudean [a,d,1]

[a] Electrical and Computer Engineering, University of British Columbia, 2332 Main Mall, Vancouver, BC V6T 1Z4, Canada
[b] The Vancouver Prostate Centre, 2660 Oak St, Vancouver, BC V6H 3Z6, Canada
[c] Department of Urologic Sciences, University of British Columbia, 2775 Laurel Street, Vancouver, BC V5Z 1M9, Canada
[d] School of Biomedical Engineering, University of British Columbia, 2222 Health Sciences Mall, Vancouver, BC V6T 1Z3, Canada
[e] Department of Pathology & Laboratory Medicine, University of British Columbia, 2211 Wesbrook Mall, Vancouver, BC V6T 1Z7, Canada

ARTICLE INFO

Dataset and code link: https://tinyurl.com/VPC-dataset, https://tinyurl.com/Zurich-dataset, https://tinyurl.com/PANDA-dataset, https://github.com/AIMLab-UBC/MS-RGCN

MSC:
41A05
41A10
65D05
65D17

Keywords:
Graph neural network
Multiple instance learning
Histopathology
Prostate cancer

ABSTRACT

Graph convolutional neural networks have shown significant potential in natural and histopathology images. However, their use has only been studied in a single magnification or multi-magnification with either homogeneous graphs or only different node types. In order to leverage the multi-magnification information and improve message passing with graph convolutional networks, we handle different embedding spaces at each magnification by introducing the Multi-Scale Relational Graph Convolutional Network (MS-RGCN) as a multiple instance learning method. We model histopathology image patches and their relation with neighboring patches and patches at other scales (i.e., magnifications) as a graph. We define separate message-passing neural networks based on node and edge types to pass the information between different magnification embedding spaces. We experiment on prostate cancer histopathology images to predict the grade groups based on the extracted features from patches. We also compare our MS-RGCN with multiple state-of-the-art methods with evaluations on several source and held-out datasets. Our method outperforms the state-of-the-art on all of the datasets and image types consisting of tissue microarrays, whole-mount slide regions, and whole-slide images. Through an ablation study, we test and show the value of the pertinent design features of the MS-RGCN.

1. Introduction

Multiple Instance Learning (MIL) is a weakly-supervised learning method that is widely utilized to train classification models with image-level annotations while detailed patch-level or pixel-level annotations are not available. Over the past years, a variety of MIL-based techniques utilizing various attention mechanisms, transformers, and Graph Neural Network (GNN) architectures have been introduced for image-based classification. Moreover, MIL-based techniques have been widely utilized in the context of giga-pixel histopathology image analysis (Zhao et al., 2020; Chen et al., 2022; Alon and Zhou, 2022; Marini et al., 2022; Adnan et al., 2020).

Computer-aided classification of histopathology images compared to natural images involves tackling several challenges as follows:

(1) The large size of the images (typically 100,000 × 100,000 pixels) makes it difficult to utilize off-the-shelf vision models. Therefore, conventional histopathology classification models rely on the extraction of smaller images, or patches, from the full image typically examined by the pathologist. Afterward, the predictions for individual patches are aggregated to achieve image-level predictions.

(2) While most existing image classification models take, as input, images at a fixed magnification, important salient morphological features at different optical zooms are in fact utilized by pathologists for accurate disease diagnosis and classification. Therefore, models are required to capture features such as cells and nuclei size, as well as high-level features such as context and tissue structure for improved classification.

(3) While classification methods require labeled data for training, acquiring detailed cell-level annotations for histopathology is known to be a complex and tedious task that requires years of training. Furthermore, the complexity of annotations results in inter-observer variability among pathologists, and even amongst experts (Nir et al., 2018).

* Corresponding author.
E-mail addresses: roozbehb@ece.ubc.ca (R. Bazargani), ali.bashashati@ubc.ca (A. Bashashati), tims@ece.ubc.ca (S. Salcudean).
[1] Co-senior authors.

https://doi.org/10.1016/j.media.2024.103197
Received 9 January 2024; Received in revised form 11 April 2024; Accepted 2 May 2024
Available online 6 May 2024

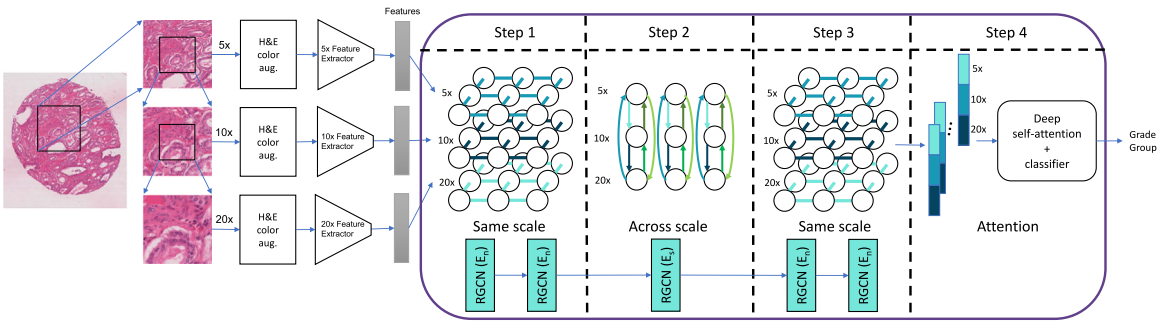

Fig. 1. Overview of the model. Patch extractions at 5×, 10×, and 20× were carried out with higher resolution patches being at the center of the prior resolution. We use a combination of stain-color and color augmentation to improve the performance and generalizability of the models to the held-out datasets. The feature extractors were trained on predicting the patch-level annotations in order to extract features. A Graph based on the patches is constructed where each node represents a patch and different edge types exist based on the relation to either neighboring or across magnification patches. Our novel method utilizes these different edge relations in four steps: (1) Two layers of RGCN have been used between neighboring nodes at each scale to get more robust features at each location by using the surrounding features; (2) One layer of the RGCN across magnification edges translates the features of each magnification to the other ones at each location; (3) Two RGCN layers on the neighboring edges aim to combine and reduce features for the final prediction, and finally; (4) A deep self-attention to better attend to the complex features and a two-layer fully-connected neural network to output the final image-level label.

(4) The performance of many existing models suffers when there is a domain shift between the training and testing data. In such scenarios, models trained on the data from one center may not perform well on data from other centers. This is widely observed in histopathology as even standard hematoxylin and eosin staining produce large color variations. Color normalization, color augmentation, and Domain Adversarial Neural Network (DANN) are among the methods that have been used to improve the generalizability of the classification methods (Macenko et al., 2009; Vahadane et al., 2016; Tellez et al., 2019a; Hashimoto et al., 2020).

To tackle the time-consuming task of performing detailed annotations by pathologists, MIL has gained substantial attention as a weakly-supervised learning method. Most of the State-Of-The-Art (SOTA) papers use attention-based MIL, either at a single magnification (Ilse et al., 2018; Lu et al., 2021) or multi-scale (Li et al., 2021; Thandiackal et al., 2022; Yao et al., 2020). The latter is done in order to better capture the analysis procedure that pathologists have and let the model look at a larger area of the image at lower magnifications while having a high resolution at higher magnifications. Since graphs efficiently describe the relation between tissue regions and patches, graph-based deep learning has shown promising results in computational histopathology in grading and survival analysis (Ahmedt-Aristizabal et al., 2021; Li et al., 2018). Zhao et al. (2020) used Deep Graph Convolutional Network (DGCN) as a MIL method at a single magnification that unveiled the potential of GNNs in histopathology. Other single magnification graph-based models have also been published in the field (Zheng et al., 2022; Chan et al., 2023; Pati et al., 2023). Furthermore, recently a few studies of multi-scale graph convolutional networks have been published. Zhang et al. (2022) utilized Graph Convolutional Networks (GCNs) at each magnification and aggregated the GCN outputs at the end for classification, and Alon and Zhou (2022) employed a multi-scale graph attentionnetwork for nuclei segmentation. Pati et al. used a hierarchical graph based on the combination of cell and tissue graphs (Pati et al., 2022). They used a homogeneous graph in a multi-scale setting to combine the information from the cells and tissue. Recently, relational and heterogeneous GNNs that utilize different edge types and assign them separate message-passing neural networks have drawn much attention (Schlichtkrull et al., 2018; Zhang et al., 2019; Wang et al., 2020; Han et al., 2021) and motivated their use in a recent model (termed H^2-MIL) that was developed for histopathology image analysis for a better representation of the relation between patches at different magnifications (Hou et al., 2022). In this work, the authors proposed a multi-scale heterogeneous graph utilizing different node types to compute resolution-aware attention for graph convolution. However, to the best of our knowledge, no previous study has explored the full potential of heterogeneous GNNs by having different node and edge types for passing information across different scales to get a better embedding space for the whole slide classification.

In this paper, we introduce the Multi-Scale Relational Graph Convolutional Network (MS-RGCN) for MIL-based aggregation of multi-scale patches in histopathology images with different node and edge types as detailed in Fig. 1. We experimented with Vision Transformers (ViT) and Convolutional Neural Network (CNN) feature extractors for the classification of prostate cancer (PCa) in Tissue Micro Arrays (TMAs) and selected the superior approach for the evaluation of the model on Whole Slide Images (WSIs). PCa classification is known to be challenging with high inter-observer classification variability amongst pathologists and therefore, suitable for evaluation of our proposed model.

Our proposed method handles the first two challenging aspects of histopathology image analysis and is evaluated under the other two as outlined earlier: (1) by passing information between patches, we are able to see a large section of the image while extracting only small patches, (2) by using different edge and node types within and between different magnifications, we preserve the various features at each magnification that might contribute to better diagnostic performance, (3) only a small part of the data is annotated and has been used to train the feature extractor from it, and (4) the models are trained on a tissue microarray (TMA) dataset from one center and tested on four diverse datasets representing biopsy and radical prostatectomy whole section slides and TMAs from Netherlands and Sweden (Bulten et al., 2022), Switzerland (Arvaniti et al., 2018), and United States.

The main contributions of this work include:

- First work in histopathology to design relational graphs with several node and edge types to tackle the problem of different embedding spaces at each magnification; showing the importance of representing pathologists' analysis procedure.
- Extensive comparison with nine SOTA models on the data from five centers representing more than 10,000 TMA cores and slides to demonstrate the importance of multi-magnification and usage of relational graphs for generalization.
- Qualitative analysis of the explainability of the model by comparing self-attention heatmaps with pathologist masks.

2. Related work

Zaheer et al. (2017) and Brendel and Bethge (2019) used bags of local features, operating on sets as a MIL task, and showed improvement on natural images. Ilse et al. (2018) was the first attention-based MIL in histopathology images. Since then, different types of MIL and aggregation of patches have been investigated on histopathology images (Tellez et al., 2019b; Yao et al., 2019, 2020; Chen et al.,

2021b; Lerousseau et al., 2021; Schirris et al., 2022; Hou et al., 2016; Carmichael et al., 2022). Moreover, multi-scale methods have been shown to improve segmentation and classification by integrating high resolution and high field-of-view from different scales (Li et al., 2021; Zhang et al., 2022; Hashimoto et al., 2020). In the following subsections, we discuss attention-based MIL methods with a specific focus on the models that utilize attention but do not have transformers in their models, Transformer-based MIL methods that explicitly used Transformers, and finally discuss the previous work in graph-based MIL.

2.1. Attention-based MIL

DeepMIL by Ilse et al. used two layers of fully connected neural networks to predict the attention (Ilse et al., 2018). Clustering-Constrained-Attention MIL (CLAM) by Lu et al. trained attention branches using clustering losses (Lu et al., 2021). Dual-Stream-MIL (DS-MIL) by Li et al. concatenated features of 5× and 20× magnifications and trained a MIL aggregator on the concatenated feature vectors (Li et al., 2021). ZoomMIL by Thandiackal et al. used attention at each magnification then summed the aggregation from each magnification and passed the results to a classifier (Thandiackal et al., 2022).

2.2. Transformer-based MIL

TransMIL by Shao et al. (2021) uses a transformer-based correlated MIL based on Nystromformer that has a nystrom-based self-attention (Xiong et al., 2021). Hierarchical Image Pyramid Transformer (HIPT) by Chen et al. (2022) uses three Vision Transformers on WSIs, ViT_{256-16}, $ViT_{4096-256}$, and $ViT_{WSI-4096}$, separately with self-supervised tasks in ViT_{256-16} and $ViT_{4096-256}$, where ViT_{b-a} denotes a ViT that takes $a \times a$ pixel images and outputs the final feature vector of $b \times b$ pixel image after combining its features. This approach used cell, patch, region, and whole image information at each of the 16, 256, 4096, and WSI levels, respectively.

2.3. Graph-based MIL

DGCN (Zhao et al., 2020), GTP (Zheng et al., 2022), HEAT (Chan et al., 2023), NAGCN (Guan et al., 2022), and WholeSIGHT (Pati et al., 2023) build their graphs based on patches from a single magnification. In particular, DGCN used a Variational AutoEncoder and Generative Adversarial Network (VAE-GAN) as a feature extractor, with three GCNs with a Self-Attention Graph Pooling (SAGPooling) (Lee et al., 2019) and two fully-connected layers as a MIL network at the end. They used two layers of fully-connected neural networks as a message passer in their three GCN layers. Moreover, they added edges between two patches if their distance was less than half of the maximum distance between patches.

A Multi-Scale Graph Wavelet Neural Network (MS-GWNN) by Zhang et al. (2022) trained a late-fusion aggregation of multi-scale GNNs where they had a graph for each magnification separately and then summed the predictions of the graphs at the end. H^2-MIL by Hou et al. (2022) utilizes heterogenous graph with different types of nodes to perform resolution-aware attention convolution.

3. Material and methods

We chose to carry out our experiments on prostate cancer. PCa is the second most common cancer in men (Siegel et al., 2019) and the fifth leading cause of death worldwide (Rawla, 2019). PCa is a heterogeneous disease, with a diverse range of dissimilar histological patterns with the same severity score, making their classification and risk stratification challenging (Nir et al., 2018). This is demonstrated by the large inter-observer variability among pathologists with overall unweighted kappa (κ) coefficient of 0.435 for general pathologists (Allsbrook et al., 2001b) and overall weighted κ range of 0.56–0.70 for urologic pathologists (Allsbrook et al., 2001a). This leads to under- or over-treatment of patients and thus impacts their survival rate, quality of life, and healthcare system costs (Berney et al., 2014).

Table 1
Data distribution in the datasets.

Dataset	Type	Train	Test	Benign	GG1	GG2	GG3	GG4	GG5	Total
Vancouver	TMA	✓	✓	221	320	217	126	131	67	1082
Zurich	TMA		✓	115	277	85	50	221	138	886
Colorado	WMS		✓	276	311	106	75	262	86	1116
Karolinska	WSI	✓		1924	1810	666	317	480	250	5447
Radboud	WSI		✓	948	801	673	908	764	963	5057

Histological patterns and/or sub-patterns are characteristic of particular tumors or groups of PCa tumors (Dive et al., 2014). The severity of morphological changes seen in the Hematoxylin and Eosin (H&E) stained tissue is graded as Gleason Patterns (GP), with severity ranging from 1 to 5. The Gleason score (GS) is based on the most and second most dominant patterns. Based on GS, there are 5 groups, in which group 2 is considered low risk, and groups 4, and 5 are considered high-grade cancerous tissue (Pierorazio et al., 2013). Epstein et al. (2016) introduced 5 International Society of Urological Pathologists (ISUP) Grade Groups (GG) based on GSs as follows: (GG1) GS $\leq$ 6, (GG2) GS3+4 = 7, (GG3) GS4+3 = 7, (GG4) GS8, and (GG5) GS9 or GS10.

We used prostate cancer datasets from different centers to evaluate our proposed model on TMAs (from Canada and Switzerland), Whole Mount Slide (WMS) regions (from the United States), and WSI biopsy datasets (from Netherlands and Sweden). The dataset from Canada consists of 1082 TMA cores from 493 radical prostatectomy patients from the Vancouver Prostate Center that were digitized at 40× magnification using an SCN400 Slide Scanner (Leica Microsystems, Wetzlar, Germany). As shown in Fig. 2, a subset of 333 TMA cores (231 patients) from the Vancouver dataset was annotated by six pathologists with different levels of experience. This subset is used for training feature extractors for all of the experiments. The majority vote was used to calculate the final annotation mask (Nir et al., 2018; Karimi et al., 2019). We used another TMA dataset from Zurich (Switzerland) to test our model. This dataset consists of 886 TMA cores at 40× magnification from 886 patients (Arvaniti et al., 2018). We also used a private dataset from the University of Colorado School of Medicine (United States) for testing, consisting of 230 WMS from 56 patients who underwent radical prostatectomy. The WMSs were digitized at 20× magnification using an Aperio ScanScope XT (Leica Biosystems, Vista, CA, USA). The Gleason patterns were annotated by the consensus of five pathologists. We randomly extracted 1116 regions from WMSs of approximately 2560 × 2560 pixels for classification and labeled them with an ISUP GG based on the GP maps in that region. Finally, to evaluate the performance of our model on WSIs and biopsy material, we utilized the publicly available portion of the PANDA challenge dataset (Bulten et al., 2022) digitized at 20× magnification. This dataset consists of two centers, Radboud (Netherlands) and Karolinska (Sweden). Table 1 shows the ISUP GG label distribution in the datasets.

3.1. Patch extraction

For patch extraction, we extracted 256 × 256 patches in a non-overlapping manner at the highest resolution, in our case 20×. At the lower magnifications (10×, and 5×), we extracted 256 × 256 patches in a way that the respective patch in the higher resolution would be in the center as demonstrated in Fig. 1. Therefore, the lower resolution patches might go out of the image bounds at the edges of the Tissue Micro Array (TMA) image where we have padded them with white pixels (value of 255 in all of the three channels). To construct the graph, the set of vertices V is computed as $V = V_{5\times} \cup V_{10\times} \cup V_{20\times}$ where $V_{m\times}$ is the set of patches from the whole image at magnification m. The

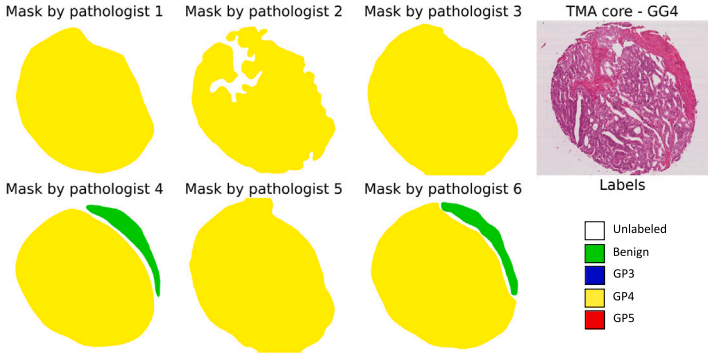

Fig. 2. Annotation of a TMA core by six pathologists.

set of edges is $E = E_n \cup E_s$ where E_n represents edges between the neighboring patches (up, down, left, and right at one magnification) and E_s represents edges between the same location patches across magnifications. We used the neighboring patches to locally improve the extracted features. Finally, we define the graph of the whole image as $G = (V, E)$. In Fig. 1, the manner of constructing a graph for the whole image is demonstrated.

3.2. Feature extractor

We compared ViT (Dosovitskiy et al., 2020) and CNN-based feature extractors in our experiments with TMAs to show the capability of our model with both feature extractor types. A detailed discussion on the comparison is available in Section *0.4* of the supplemental material. However, for the remainder of our experiments, we utilized the CNN-based feature extractor as it provided superior results.

Furthermore, we trained the feature extractor on the Vancouver dataset with their patch-level Gleason Score annotations (Nir et al., 2018; Karimi et al., 2019). For the CNN model, we selected ResNet18 (He et al., 2016) as it was utilized in the previous studies on the Vancouver and Colorado datasets (Bazargani et al., 2023). Its smaller number of parameters compared to the other CNN models provided us the opportunity to update all the weights of the pre-trained ResNet on the ImageNet dataset (Russakovsky et al., 2015) using transfer-learning by performing patch-based classification. We used the weighted cross entropy loss function to tackle the imbalanced classes in the datasets. Furthermore, to improve model generalizability on held-out sets, we utilize a combination of color normalization and augmentation (Bazargani et al., 2023), that was inspired by Tellez et al. (2019a) and Boschman et al. (2022).

3.3. Graph-based MIL

In contrast to the previous studies based on graph neural networks within the context of histopathology, we used different type of edges between magnifications in order to provide the best early fusion of information between different magnifications. This follows Dwivedi et al. (2022), which showed that early fusion of information helped the model perform better than late fusion.

The main difficulty in combining features from different magnifications is that morphological features at different magnifications have different interpretations. To solve this issue, after constructing our graph, we gave each node a type based on the magnification the node belongs to. So the set of types T that a vertex (node) $v \in V$ belongs to consists of three elements $T = \{5, 10, 20\}$. We used different edge types for connecting different types of nodes. The set of types R that an edge e belongs to is the Cartesian product of T with itself

$$R = T \times T = \{(5,5),(10,10),(20,20),(5,10),(10,5),(5,20),(20,5),(10,20),(20,10)\} \quad (1)$$

where (a, b) indicates the type of edge that goes from a vertex of type a to a vertex of type b. We partition the set of edge types into neighboring edge types, and scaling edge types as $R_n = \{(5,5),(10,10),(20,20)\}$ on E_n and $R_s = \{(5,10),(10,5),(5,20),(20,5),(10,20),(20,10)\}$ on E_s, respectively, where $R_n \cap R_s = \emptyset$ and $R_n \cup R_s = R$. As illustrated in Fig. 1, our method consists of four sections based on the sets R_n and R_s, as discussed next.

We utilized two same-scale layers so that each node can get the information from neighbors with a distance of two or less (13 nodes in total) while with one layer, we would have access to the information of fewer patches (four in this case). Furthermore, increasing the number of layers (i.e., more than two) would lead to the incorporation of a large number of patches and over-smoothing which is a known issue in graph neural networks. In addition, this could be inefficient from the computational perspective. The purpose of adding a cross-scale layer is to pass information across magnifications. As all patches across magnifications are connected, there is no need to have 2 cross-scale layers. As for their order, we started with same-scale layers for the following reasons:

- We merge the information from 13 patches in comparison to 3 for cross-scales.
- The embedding spaces from the same-scale nodes are the same, making it easier to use low-pass filter GNNs to obtain robust features at the same scale, and then communicate the information across scales.

At the first step, we used two layers of RGCN (Schlichtkrull et al., 2018), based on Graph Convolutional Networks (GCN) (Kipf and Welling, 2016). In a RGCN, we have

$$h_i' = b + W_{root} \cdot h_i + \frac{1}{|R_k|} \sum_{r \in R_k} \sum_{j \in N_r(i)} \frac{1}{|N_r(i)|} W_r \cdot h_j, \quad (2)$$

where h_i, and h_i' are the current and output embeddings of node i, b is the bias vector, R_k is the edge type, which can be either R_n or R_s, $N_r(i)$ is the set of neighboring nodes of node i based on the edges of type r, and W_{root} and W_r are trainable weights that are applied to the node itself and neighboring nodes based on edge type r, respectively. We applied these layers on the R_n edges to make the extracted features more robust by a learned weighted averaging based on the features of neighboring patches. This will let the network see features of neighboring patches at the distance of 2 vertices in the graph which are 12 vertices. We did not use non-linear activation functions here, as we wanted linear mapping of the features that we already have in order to see the greater distance without changing the features. To experiment the benefit, in the ablation study 4.5 there is a case where we have used a non-linear activation function in this step.

Now, having more robust features at each scale, we interpret the features of other scales to the features of the scale where the node belongs to in the message passing network. For this purpose, we used one layer of RGCN on R_s edges with a layer-normalization (Ba et al.,

2016) and Rectified Linear Unit (ReLU) activation function. Different edge types and as a result, different message passing networks will let the model effectively pass the information between different node types. For instance, the network on edge type $(10, 20)$ learns how to translate the features at 10× to features at 20×. An example of how a node at 20× will receive the information from other magnifications is based on Eq. (2). Information from 5× and 10× will be translated to 20× via $(5, 20)$ and $(10, 20)$ edges, respectively. Then, the aggregation of the translated information and the information from the 20× vertex will replace the node features. This multi-scale idea lets each node have a broad field of view at a higher resolution. For example, a node at 20× can see 13 patches at 5× and vice versa. Moreover, it is important to note that at lower magnifications it is much easier to distinguish benign from cancerous regions, and at higher magnifications it is easier to distinguish the subtypes from each other.

Next, we start to change the features in each magnification using two layers of RGCN on R_n edges with layer-normalization and ReLU activation function. It is important to note that each magnification has different features so we cannot combine them with GCN layers instead of RGCN layers. This is why after RGCN layers we concatenate the features of different scale nodes at each location instead of averaging them. Next, we carry out our final step which is pooling this bag of nodes.

Finally, each location in the main image has a concatenated feature vector. Considering the complexity of the histopathology images and their features, we use the pooling operation from Ilse et al. (2018), where the attention values are learned by a neural network. After pooling, a two-layer fully-connected classifier was trained to recognize the image-level label.

4. Results and discussion

4.1. Experimental setup

We used 5-fold cross-validation training to eliminate selection bias and split the TMA cores into training, validation, and test sets based on patients, regardless of the number of cores available per patient and slides in order to have no overlap in feature extractor and MIL methods training. We used two different criteria: macro-averaged Area Under the Receiver Operating Characteristic Curve (AUC), and quadratic-weighted Cohen's κ (Cohen, 1960) to compare the performance of the models. Cohen's κ has been used on histopathology to determine the agreement among pathologists and artificial intelligence models based on the confusion matrix (Ström et al., 2020). For computing quadratic-weighted-κ, we weigh diagonal elements of the confusion matrix 0, and weigh the rest of the elements based on its quadratic distance. As a result, we would penalize the prediction of ground truth of GG2 as GG5 more than predicting it as GG3. AUC can take a value between 0 and 1 where 0.5 means random prediction, and κ can take a value ranging from −1 to 1 where between −1 and 0 the agreement is by chance and 1 means complete agreement. Thus, higher AUC and higher κ are better.

4.1.1. Feature extraction setup

The feature extractor at each magnification of our proposed workflow was trained with a batch size of 32 patches of size 256 × 256. The patches were extracted at the actual magnification of the image (40× for Vancouver and Zurich and 20× for the rest of the datasets) and downsampled to the required magnification for the feature extractor. The classes consist of Background, Benign, GP3, GP4, GP5, and Others that includes stroma or non-tumor regions. We used the Adam optimizer (Kingma and Ba, 2014) with learning rate of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. In the end, we extracted embeddings of each patch at the average pooling layer of ResNet18. Note that we trained the feature extractor only on the Vancouver dataset. We also trained separate MIL models for TMAs, WMS regions, and WSIs. The MIL method for TMAs and WMS regions were trained based on the Vancouver dataset. For WSIs, we utilized the Karolinska dataset for MIL training and tested it on Radboud as the class distribution for the Radboud dataset is more uniform and has GP annotations for model evaluations. WSIs are larger and have more patches than TMAs and therefore necessitated the training of another MIL for these datasets.

4.1.2. MIL setup

To compare the different MIL methods, we used the same feature extractor and froze its weights in all of the experiments. The task for training MILs was to classify the whole image (TMA, WMS, or WSI) into 6 classes of Benign (BN), and ISUP GG1-5. The batch size of 32 images and an Adam optimizer with a learning rate of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ were used for training. The scheduler of the learning rate was *ReduceLROnPlateau* with factor = 0.3, and patience = 10.

We also compared the models with the ResNet18 feature extractor on the Colorado and PANDA datasets. As the slide-to-patients mapping of the PANDA challenge, comprising the Radboud and Karolinska datasets, is not published, we opted to train on one dataset and test on the other one to make sure that we have split the patients (as opposed to the patches or slides) in our training and test set. Therefore, we trained the MIL methods on the Karolinska dataset and tested them on the Radboud dataset. We chose to test on the Radboud dataset because it is more balanced compared to the Karolinska dataset, and because it has Gleason score-based annotations while the Karolinska dataset does not. We have also tested training on the Radboud dataset and testing on the Karolinska and noted that the overall performance of the models is much lower.

4.2. Comparison with state-of-the-art

We compared our method with SOTA models in single and multiple magnifications. For single magnification models, we evaluated the aggregators of DeepMIL, CLAM-SB, DGCN, HIPT-MIL, and TransMIL. DeepMIL and CLAM-SB are attention-based, DGCN is a graph-based, and HIPT-MIL and TransMIL are transformer-based representative models. For multi-magnification models, we compared our model with DS-MIL, and ZoomMIL as attention-based and MS-GWNN and H^2-MIL as a Graph-based MIL model.

The results of comparing our model with SOTA models are shown in Table 2. It can be seen that MS-RGCN is outperforming all of the models . MS-RGCN improves SOTA models performance by up to 12.4% in quadratic weighted Cohen's κ, and 4.56% in macro averaged AUC on the Vancouver set and up to 10.2% in κ and 6.43% in AUC on the Zurich set. We can see that DeepMIL has the closest performance to MS-RGCN on the source dataset. However, its performance drops significantly on the Zurich dataset where we have a large gap of 4.7% in κ and 2.60% in AUC. Furthermore, ZoomMIL is the best SOTA method in terms of generalizability, however MS-RGCN performs better by a large margin on the Vancouver and Zurich datasets.

It is important to note that, for a fair comparison of our proposed model with the state-of-the-art, we tried various hyper-parameter settings for competing algorithms to make sure hyper-parameter setting did not contribute to inferior performance of the state-of-the-art algorithms (further details in Supplemental text; Section *0.2.*).

Results in Table 2 indicate that MS-RGCN outperforms the best SOTA models by 6.4%, and 2.7% in κ and 1.61% and 2.50% AUC, in Colorado and Radboud datasets, respectively. Thus, the proposed MS-RGCN is outperforming nine SOTA methods in various datasets and feature extractors. It is worth mentioning that the lower performance on the Radboud dataset from the PANDA challenge in comparison to the reported results in that challenge (Bulten et al., 2022) is due to the difference in the test datasets. Since the PANDA challenge *test* dataset is not publicly available, we utilized the two training datasets that were available. In their paper (Bulten et al., 2022) and data website, it is mentioned that due to the high number of slides in the training set, the authors relied on pathology reports to extract Gleason scores associated with patients in the training data. However, these scores could be noisy and less reliable in comparison to the validation and test sets where the slides were annotated by a consensus of several pathologists.

Table 2
Comparison with state-of-the-art MIL methods with ResNet18 feature extractors. We report average and standard deviation of macro-averaged AUC and quadratic weighted κ on 5-folds. The best performance is shown in **bold**.

Method	Vancouver dataset		Zurich dataset		Colorado dataset		Radboud dataset	
	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)
DeepMIL (Ilse et al., 2018)	0.714 ± 0.019	84.39 ± 1.44	0.734 ± 0.029	82.64 ± 1.59	0.580 ± 0.052	78.84 ± 0.78	0.388 ± 0.075	68.60 ± 4.92
CLAM-SB (Lu et al., 2021)	0.672 ± 0.050	84.11 ± 1.65	0.735 ± 0.007	82.18 ± 0.95	0.570 ± 0.047	78.67 ± 1.72	0.420 ± 0.070	70.84 ± 4.18
DGCN (Zhao et al., 2020)	0.640 ± 0.068	80.62 ± 1.45	0.638 ± 0.095	78.81 ± 1.63	0.561 ± 0.035	75.45 ± 1.35	0.412 ± 0.089	69.76 ± 2.82
HIPT-MIL (Chen et al., 2022)	0.666 ± 0.044	83.52 ± 1.37	0.724 ± 0.013	81.56 ± 1.20	0.576 ± 0.033	78.47 ± 2.08	0.457 ± 0.060	71.14 ± 2.32
TransMIL (Shao et al., 2021)	0.603 ± 0.086	81.75 ± 1.74	0.679 ± 0.013	81.41 ± 1.17	0.514 ± 0.117	76.19 ± 4.31	0.366 ± 0.052	69.89 ± 2.90
DS-MIL (Li et al., 2021)	0.618 ± 0.049	80.72 ± 1.71	0.710 ± 0.040	81.93 ± 2.02	0.597 ± 0.034	78.13 ± 0.61	0.531 ± 0.036	76.34 ± 2.57
ZoomMIL (Thandiackal et al., 2022)	0.689 ± 0.065	81.76 ± 1.41	0.738 ± 0.022	83.62 ± 0.87	0.601 ± 0.031	78.99 ± 1.54	0.550 ± 0.040	75.94 ± 1.86
MS-GWNN (Zhang et al., 2022)	0.612 ± 0.101	82.55 ± 1.17	0.735 ± 0.024	82.16 ± 1.23	0.604 ± 0.034	79.25 ± 1.81	0.374 ± 0.102	74.91 ± 1.35
H^2-MIL (Hou et al., 2022)	0.676 ± 0.094	82.29 ± 2.59	0.739 ± 0.043	80.99 ± 2.11	0.503 ± 0.094	76.14 ± 3.75	0.479 ± 0.065	75.77 ± 2.47
MS-RGCN (ours)	$\mathbf{0.727 \pm 0.028}$	$\mathbf{85.18 \pm 1.45}$	$\mathbf{0.781 \pm 0.013}$	$\mathbf{85.24 \pm 1.07}$	$\mathbf{0.668 \pm 0.031}$	$\mathbf{80.86 \pm 0.89}$	$\mathbf{0.577 \pm 0.025}$	$\mathbf{78.84 \pm 2.73}$

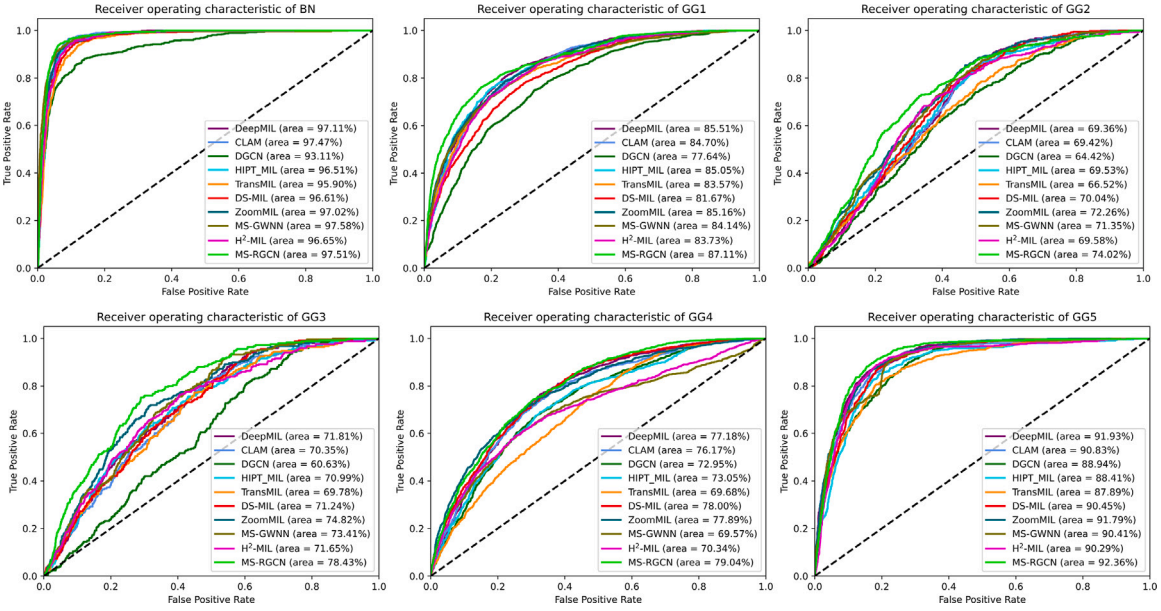

Fig. 3. AUC performance of SOTA models on the Zurich (held-out) dataset based on each class across all of the folds. It can be seen that our model is the best in all of the ISUP GGs except benign which is the easiest task and our model is the second best with an insignificant difference of 0.07%.

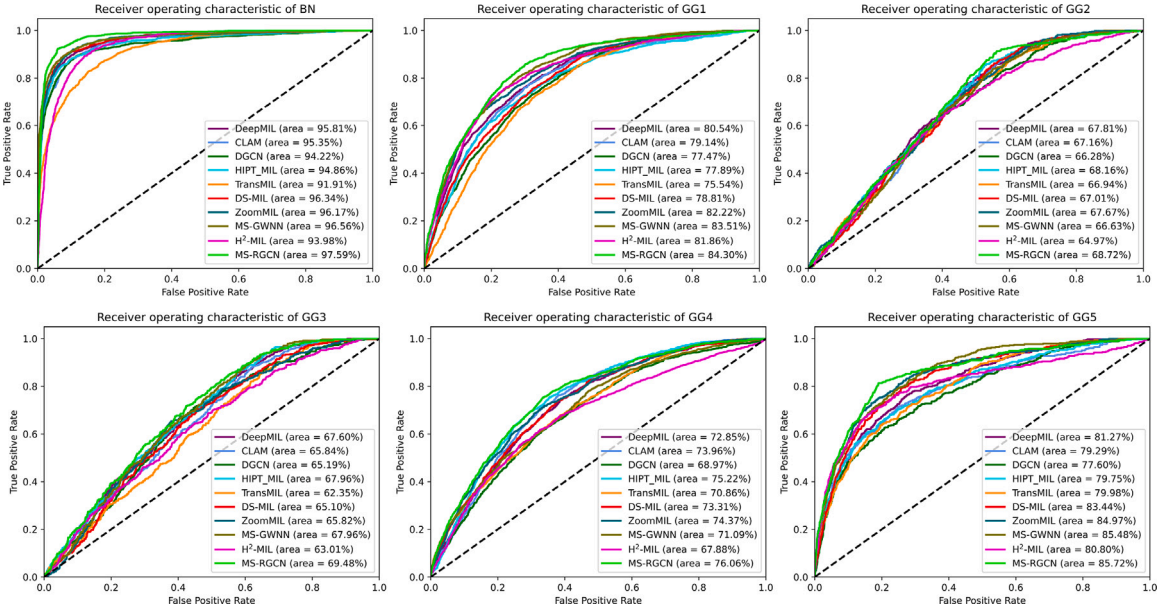

Fig. 4. AUC performance of SOTA models on the Colorado dataset based on each class across all of the folds. Our model is the best in all of the ISUP GGs.

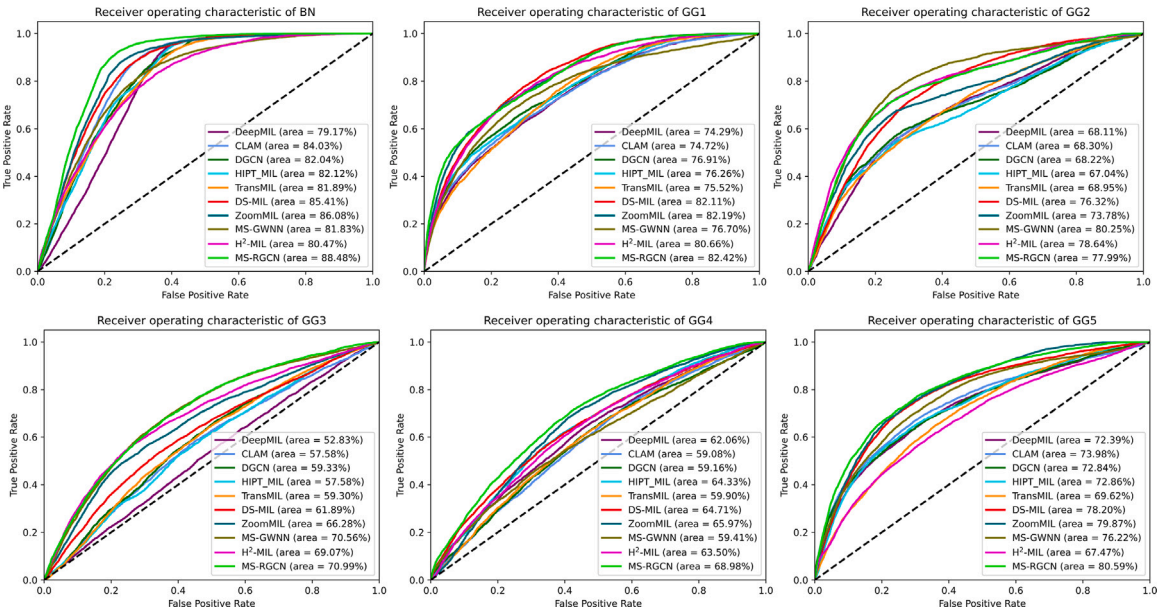

Fig. 5. AUC performance of SOTA models on the Radboud dataset based on each class across all of the folds. Our model is the best in all of the ISUP GGs.

4.3. Class-based performance

In Fig. 3, the average performance of state-of-the-art models for each class of the Zurich dataset is shown. Our model is performing well across all of the classes, being the first in all of GG1-5 with a margin of 1.6%, 1.76%, 3.61%, 1.15%, and 0.43% in one-vs-all AUC compared to the best SOTA model, respectively, and second in benign with negligible difference of 97.58%–97.51% = 0.07% after MS-GWNN. Moreover, it performs substantially better GG2 (GS3+4) and GG3 (GS4+3) classification tasks which are deemed to be more challenging where the model has to distinguish GP3 and GP4, by up to 9.60% and 17.80% in one-vs-all AUC, respectively.

In addition, the average performance of SOTA models for each class of the Colorado and Radboud datasets is shown in Figs. 4 and 5. In the Colorado dataset, it can be seen that MS-RGCN outperforms SOTA models in all ISUP GG classification tasks. Improvement of the MS-RGCN in comparison to the best SOTA model in each GG is 1.03%, 0.79%, 0.56%, 1.52%, 0.84%, and 0.24% in AUC. In the Radboud dataset, the proposed model outperforms SOTA models in all of the GGs except for GG2 where it is the third-best model after MS-GWNN and H^2-MIL.

4.4. Visualization of the prediction heatmaps

Fig. 6 illustrates the attention heatmap of the model for the final prediction where red shows the highest attention and blue the lowest. We have plotted the attention for one sample per class and the model attention demonstrates the ability of the model to select the important regions for the final prediction. In the first column, we see a benign core, where the model has focused on the benign gland structure to identify the image as benign. In the second column, the model has focused on the whole core which contains GP3. In the third and fourth columns, the model has attended to both GP3 and GP4 regions to correctly classify the cores to GG2 (GS3+4) and GG3 (GS4+3). In the fifth column, attentions are only on the annotated regions (GP4) and the model has successfully ignored the unlabeled regions. Finally, in the last column, although we have GP3-5, the model has focused on GP5 to correctly predict the final ISUP GG.

In addition, Fig. 7 shows the heatmaps on WSIs of the Radboud dataset. It can be observed that for the benign slide, the model has focused on the Benign section. In GG1, it ignored benign regions and only focused on the GP3 regions. In GG2 and GG3 slides, it focused more on the GP4 section which is more aggressive but also paid attention to the GP3 section to correctly classify them as GP3+GP4 (GG2) and GP4+GP3 (GG3). Finally, on the GG4 and GG5, attentions were on GP4 and GP4+GP5 which are the important components for predicting the grade group of the slide. Based on these results, we can also observe that the model does not suffer from over-smoothing. This is evidenced by the attention heatmap which is different across patches (nodes). This shows there is a difference between node embeddings across the graph that has resulted in different attention values.

4.5. Ablation study

For the ablation study, we experimented with seven models in addition to our model. MS-RGCN-5×, MS-RGCN-10×, and MS-RGCN-20× are the proposed MS-RGCN in single 5×, 10×, and 20× magnifications. MS-RGCN-GE is the ablation study on graph edge definitions where there is an edge between two patches in one magnification (E_n) if their distance is less than half of the maximum distance, similar to the DGCN (Zhao et al., 2020) edge definition. MS-RGCN without Scale Edges (MS-RGCN-w/o-SE) is obtained by removing the across-magnification edges in step 2 and only using steps 1, 3, and 4 in Fig. 1. Multi-Scale Graph Convolution Network (MS-GCN) is the same model as MS-RGCN but a homogenous graph where all edges and nodes are of the same type and the same message-passing network is used in all of the connections. In other words, MS-GCN does not have the relational and heterogeneous features that the proposed MS-RGCN has. MS-RGCN-4ReLU is the MS-RGCN model but with ReLU activation functions in the first step of the model where we did not use activation functions to prevent adding non-linearity and changing the features while we are trying to make them robust by seeing neighbors in a distance of 2 nodes in the graph.

As it can be seen in Table 3, MS-RGCN is substantially better than single magnification models, showing the effectiveness of multi-scale learning of the model. MS-RGCN-GE demonstrates that our locally defined edge is better compared to more global ones, as in DGCN. MS-RGCN-w/o-SE shows the importance of the second step in the model where we pass information across different scales. MS-GCN is very close to single magnification results, especially if we look at the κ values. This shows the importance of using a relational graph and edge types in order to not mix information from different scales and get the highest possible performance with multi-scale information. Finally,

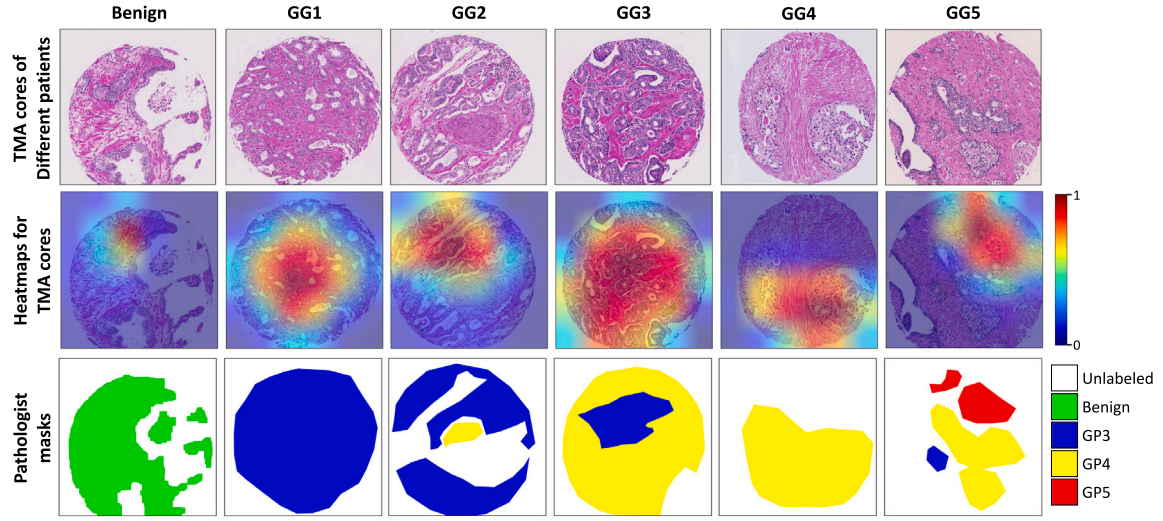

Fig. 6. Self-attention heatmaps of the MS-RGCN on one sample from each class in the Zurich dataset. It can be seen that the model has focused on the important location for the classification, such as where the gland is, where the annotations are, and where the most important annotation is.

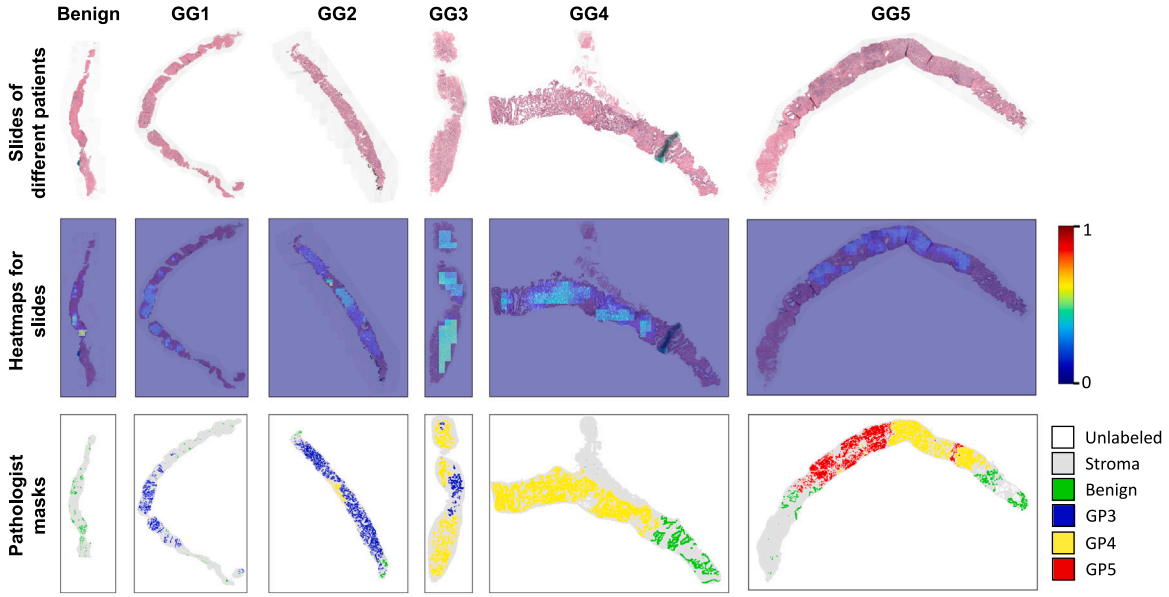

Fig. 7. Self-attention heatmaps of the MS-RGCN on one sample from each class in the Radboud dataset. It can be seen that the model has focused on the important location for the classification, such as where the gland is, where the annotations are, and where the most important annotation is.

Table 3
Ablation study on MIL methods with ResNet18 feature extractor. We report the average and standard deviation of macro-averaged AUC and quadratic weighted κ on 5-folds. The best performance is shown in **bold**.

Method	Vancouver dataset		Zurich dataset		Colorado dataset		Radboud dataset	
	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)
MS-RGCN-5×	0.623 ± 0.028	80.62 ± 0.31	0.719 ± 0.020	82.75 ± 1.44	0.604 ± 0.036	78.11 ± 0.59	0.481 ± 0.038	77.37 ± 1.39
MS-RGCN-10×	0.716 ± 0.022	83.35 ± 1.03	0.758 ± 0.047	83.66 ± 2.29	0.555 ± 0.103	79.25 ± 1.39	0.545 ± 0.044	77.79 ± 1.39
MS-RGCN-20×	0.704 ± 0.040	83.01 ± 2.04	0.721 ± 0.047	81.53 ± 1.62	0.588 ± 0.083	76.40 ± 4.02	0.386 ± 0.061	68.84 ± 3.72
MS-RGCN-GE	0.691 ± 0.034	83.79 ± 1.25	0.757 ± 0.015	83.99 ± 1.52	0.656 ± 0.017	80.16 ± 1.57	Out of memory error	
MS-RGCN-w/o-SE	0.708 ± 0.034	84.20 ± 0.82	0.756 ± 0.042	84.31 ± 1.40	0.649 ± 0.024	78.93 ± 1.59	0.565 ± 0.019	77.30 ± 3.42
MS-GCN	0.694 ± 0.037	83.49 ± 1.84	0.754 ± 0.061	84.08 ± 2.09	0.657 ± 0.036	78.88 ± 1.40	0.560 ± 0.030	77.35 ± 3.28
MS-RGCN-4ReLU	0.718 ± 0.044	84.38 ± 1.77	0.779 ± 0.039	84.16 ± 1.99	0.649 ± 0.047	79.83 ± 1.16	0.571 ± 0.033	78.18 ± 2.67
MS-RGCN (ours)	**0.727 ± 0.028**	**85.18 ± 1.45**	**0.781 ± 0.013**	**85.24 ± 1.07**	**0.668 ± 0.031**	**80.86 ± 0.89**	**0.577 ± 0.025**	**78.84 ± 2.73**

Table 4
Comparing Inference Time (IT) for TMA (Zurich dataset) and WSI (Radboud dataset) in milliseconds and number of parameters in millions with state-of-the-art MIL methods. The best performance is shown in **bold**.

Method	IT (TMA)	IT (WSI)	Number of parameters
DeepMIL	17.6	**83.3**	$\mathbf{0.13 \times 10^6}$
CLAM-SB	**17.2**	88.2	0.53×10^6
DGCN	21.4	89.3	0.66×10^6
HIPT-MIL	20.4	94.2	0.66×10^6
TransMIL	27.3	90.4	2.41×10^6
DS-MIL	17.6	86.6	1.22×10^6
ZoomMIL	20.2	86.9	3.17×10^6
MS-GWNN	17.4	86.4	0.21×10^6
H^2-MIL	19.8	96.5	1.19×10^6
MS-RGCN (ours)	**17.2**	86.5	4.24×10^6

MS-RGCN-4ReLU shows that seeing a larger area without changing the embeddings with non-linear activation functions is a better choice compared to having non-linear activation functions.

4.6. Inference time and number of parameters

Table 4 illustrates the inference time and the number of parameters for each MIL model. It can be observed that multi-scale methods have more parameters and at the same time gain better performance since they mimic pathologists' analysis better. For instance, ZoomMIL, the best SOTA model on Zurich and Colorado datasets, has the highest number of parameters (3.17 million) and, similarly, our proposed MS-RGCN has more parameters but also has superior performance compared to other techniques. In addition, our proposed method and CLAM-SB have the lowest inference time of 17.2 ms in TMAs and our inference time is close to the best model in WSIs. The models are trained and tested using an NVIDIA® Tesla® V100 GPU with 16 GB memory.

It is worth mentioning that without parallelization, single-magnification models would be three times faster in feature extraction. However, in our case, since ResNet18 is small we are able to handle the three magnifications in parallel, resulting in the same feature extraction time. Thus, having the same inference time in the feature extraction part of all of the models, we did not report that time and only focused on the MIL inference time. For the reported ZoomMIL performance we did not ignore any patches, as it achieved the highest performance this way (a trade-off between performance and time), resulting in having equal feature extraction time as ours.

5. Conclusion

We believe our work is an important step towards improving GCN performance on different types of information for multiple instance learning with the help of RGCNs and different edge types. To the best of our knowledge, this is the first paper that uses GCNs in multi-scale analysis of histopathology images to their fullest potential by using a combination of different edge types between neighboring patches and patches at different scales. This is helpful in handling different feature types in MIL problems. The proposed method outperforms SOTA on the source and held-out datasets and is capable of classifying complex classes substantially better. We developed this model for PCa which is a heterogeneous disease with a diverse range of histological patterns which make it challenging for diagnosis and prognosis. PCa remains the third-leading cause of cancer death in men, along with prostate-specific antigen, histopathology analysis of biopsy samples is a crucial step in deciding patient treatment (Litwin and Tan, 2017).

The limitation of this work is that in the end, attention-based mechanisms are required for image-level prediction in our model. Future directions include dedicating a node (as an output) to combine information from different levels of RGCNs; this would be similar to having skip connections for the final prediction. In addition, we only focused on the patch-based graphs as cell graphs (depending on the architecture) may ignore important prognostic tissue features such as stroma and only focus on the small image regions (Chen et al., 2021a). An interesting direction to explore is combining our work, comprising the use of different node and edge types, with the approach from Pati et al. (2022), comprising both cell and tissue graphs. We would like to point out that the idea of handling different embedding spaces using RGCNs could motivate multi-modal problems, similar to Dwivedi and Shao et al. work (Dwivedi et al., 2022; Shao et al., 2020).

CRediT authorship contribution statement

Roozbeh Bazargani: Writing – original draft, Validation, Methodology, Formal analysis. **Ladan Fazli:** Supervision, Data curation. **Martin Gleave:** Supervision, Data curation. **Larry Goldenberg:** Supervision, Data curation. **Ali Bashashati:** Writing – review & editing, Supervision. **Septimiu Salcudean:** Writing – review & editing, Supervision.

Data and code availability

Majority of the data associated with this manuscript is publicly available as follows:

(a) Vancouver dataset as part of the Gleason 2019 Challenge https://tinyurl.com/VPC-dataset,
(b) Zurich dataset: https://tinyurl.com/Zurich-dataset,
(c) Karolinska and Radboud datasets from PANDA challenge: https://tinyurl.com/PANDA-dataset. The code associated with this work is available on https://github.com/AIMLab-UBC/MS-RGCN.

Acknowledgments

This work was supported by the Canadian Institute of Health Research (CIHR) [Grant # 450849] and Michael Smith Health Research Scholar Award for AB.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103197.

References

Adnan, M., Kalra, S., Tizhoosh, H.R., 2020. Representation learning of histopathology images using graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 988–989.

Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., Petersson, L., 2021. A survey on graph-based deep learning for computational histopathology. Comput. Med. Imaging Graph. 102027.

Allsbrook, Jr., W.C., Mangold, K.A., Johnson, M.H., Lane, R.B., Lane, C.G., Amin, M.B., Bostwick, D.G., Humphrey, P.A., Jones, E.C., Reuter, V.E., et al., 2001a. Interobserver reproducibility of gleason grading of prostatic carcinoma: urologic pathologists. Hum. Pathol. 32 (1), 74–80.

Allsbrook, Jr., W.C., Mangold, K.A., Johnson, M.H., Lane, R.B., Lane, C.G., Epstein, J.I., 2001b. Interobserver reproducibility of gleason grading of prostatic carcinoma: general pathologist. Hum. Pathol. 32 (1), 81–88.

Alon, Y., Zhou, H., 2022. Neuroplastic graph attention networks for nuclei segmentation in histopathology images. arXiv preprint arXiv:2201.03669.

Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M., 2018. Automated gleason grading of prostate cancer tissue microarrays via deep learning. Sci. Rep. 8 (1), 1–11.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Bazargani, R., Chen, W., Sadeghian, S., Asadi, M., Boschman, J., Darbandsari, A., Bashashati, A., Salcudean, S., 2023. A novel h and e color augmentation for domain invariance classification of unannotated histopathology prostate cancer images. In: Medical Imaging 2023: Digital and Computational Pathology. Vol. 12471, SPIE, pp. 224–229.

Berney, D.M., Algaba, F., Camparo, P., Compérat, E., Griffiths, D., Kristiansen, G., Lopez-Beltran, A., Montironi, R., Varma, M., Egevad, L., 2014. The reasons behind variation in Gleason grading of prostatic biopsies: areas of agreement and misconception among 266 European pathologists. Histopathology 64 (3), 405–411.

Boschman, J., Farahani, H., Darbandsari, A., Ahmadvand, P., Van Spankeren, A., Farnell, D., Levine, A.B., Naso, J.R., Churg, A., Jones, S.J., et al., 2022. The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images. J. Pathol. 256 (1), 15–24.

Brendel, W., Bethge, M., 2019. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. arXiv preprint arXiv:1904.00760.

Bulten, W., Kartasalo, K., Chen, P.-H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., et al., 2022. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. Nat. Med. 28 (1), 154–163.

Carmichael, I., Song, A.H., Chen, R.J., Williamson, D.F., Chen, T.Y., Mahmood, F., 2022. Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 387–397.

Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L., 2023. Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15661–15670.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155.

Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., 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–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. Springer, pp. 339–349.

Chen, R.J., Lu, M.Y., Weng, W.-H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., 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. pp. 4015–4025.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Dive, A.M., Bodhade, A.S., Mishra, M.S., Upadhyaya, N., 2014. Histological patterns of head and neck tumors: An insight to tumor histology. J. Oral Maxillofac. Pathol.: JOMFP 18 (1), 58.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dwivedi, C., Nofallah, S., Pouryahya, M., Iyer, J., Leidal, K., Chung, C., Watkins, T., Billin, A., Myers, R., Abel, J., et al., 2022. Multi stain graph fusion for multimodal integration in pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1835–1845.

Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., Srigley, J.R., Humphrey, P.A., 2016. The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma. Am. J. Surg. Pathol. 40 (2), 244–252.

Guan, Y., Zhang, J., Tian, K., Yang, S., Dong, P., Xiang, J., Yang, W., Huang, J., Zhang, Y., Han, X., 2022. Node-aligned graph convolutional network for whole-slide image representation and classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18813–18823.

Han, G., He, Y., Huang, S., Ma, J., Chang, S.-F., 2021. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3263–3272.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I., 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3852–3861.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2424–2433.

Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L., 2022. H^2-MIL: Exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 933–941.

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

Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E., 2019. Deep learning-based gleason grading of prostate cancer from histopathology images—Role of multiscale decision aggregation and data augmentation. IEEE J. Biomed. Health Inform. 24 (5), 1413–1426.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lee, J., Lee, I., Kang, J., 2019. Self-attention graph pooling. In: International Conference on Machine Learning. PMLR, pp. 3734–3743.

Lerousseau, M., Vakalopoulou, M., Deutsch, E., Paragios, N., 2021. SparseConvMIL: Sparse convolutional context-aware multiple instance learning for whole slide image classification. In: MICCAI Workshop on Computational Pathology. PMLR, pp. 129–139.

Li, B., Li, Y., Eliceiri, K.W., 2021. 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. pp. 14318–14328.

Li, R., Yao, J., Zhu, X., Li, Y., Huang, J., 2018. Graph CNN for survival analysis on whole slide pathological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 174–182.

Litwin, M.S., Tan, H.-J., 2017. The diagnosis and treatment of prostate cancer: a review. JAMA 317 (24), 2532–2542.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, pp. 1107–1110.

Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., van Rijthoven, M., Aswolinskiy, W., Bokhorst, J.-M., Podareanu, D., Petters, E., et al., 2022. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. NPJ Digit. Med. 5 (1), 1–18.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Wilson, R.S., et al., 2018. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Med. Image Anal. 50, 167–180.

Pati, P., Jaume, G., Ayadi, Z., Thandiackal, K., Bozorgtabar, B., Gabrani, M., Goksel, O., 2023. Weakly supervised joint whole-slide segmentation and classification in prostate cancer. arXiv preprint arXiv:2301.02933.

Pati, P., Jaume, G., Foncubierta-Rodriguez, A., Feroce, F., Anniciello, A.M., Scognamiglio, G., Brancati, N., Fiche, M., Dubruc, E., Riccio, D., et al., 2022. Hierarchical graph representations in digital pathology. Med. Image Anal. 75, 102264.

Pierorazio, P.M., Walsh, P.C., Partin, A.W., Epstein, J.I., 2013. Prognostic g leason grade grouping: data based on the modified g leason scoring system. BJU Int. 111 (5), 753–760.

Rawla, P., 2019. Epidemiology of prostate cancer. World J. Oncol. 10 (2), 63.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J., 2022. DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. Med. Image Anal. 79, 102464.

Schlichtkrull, M., Kipf, T.N., Bloem, P., Berg, R.v.d., Titov, I., Welling, M., 2018. Modeling relational data with graph convolutional networks. In: European Semantic Web Conference. Springer, pp. 593–607.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al., 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural Inf. Process. Syst. 34, 2136–2147.

Shao, Y., Nir, G., Fazli, L., Goldenberg, L., Gleave, M., Black, P., Wang, J., Salcudean, S., 2020. Improving prostate cancer classification in H&E tissue micro arrays using Ki67 and P63 histopathology. Comput. Biol. Med. 127, 104053.

Siegel, R.L., Miller, K.D., Jemal, A., 2019. Cancer statistics, 2019. CA Cancer J. Clin. 69 (1), 7–34.

Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., et al., 2020. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol. 21 (2), 222–232.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., Van Der Laak, J., 2019a. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544.

Tellez, D., Litjens, G., van der Laak, J., Ciompi, F., 2019b. Neural image compression for gigapixel histopathology image analysis. IEEE Trans. Pattern Anal. Mach. Intell. 43 (2), 567–578.

Thandiackal, K., Chen, B., Pati, P., Jaume, G., Williamson, D.F., Gabrani, M., Goksel, O., 2022. Differentiable zooming for multiple instance learning on whole-slide images. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI. Springer, pp. 699–715.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N., 2016. Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans. Med. Imaging 35 (8), 1962–1971.

Wang, H., Zheng, W.-s., Yingbiao, L., 2020. Contextual heterogeneous graph network for human-object interaction detection. In: European Conference on Computer Vision. Springer, pp. 248–264.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V., 2021. Nyströmformer: A nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35, pp. 14138–14148.

Yao, J., Zhu, X., Huang, J., 2019. Deep multi-instance learning for survival prediction from whole slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 496–504.

Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J., 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med. Image Anal. 65, 101789.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J., 2017. Deep sets. Adv. Neural Inf. Process. Syst. 30.

Zhang, M., Dong, B., Li, Q., 2022. MS-GWNN: multi-scale graph wavelet neural network for breast cancer diagnosis. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Zhang, C., Song, D., Huang, C., Swami, A., Chawla, N.V., 2019. Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 793–803.

Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B., Fan, X., et al., 2020. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4837–4846.

Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B., 2022. A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging 41 (11), 3003–3015.