Towards Explainable Graph Representations in Digital Pathology

Guillaume Jaume* [1 2] **Pushpak Pati*** [1 3] **Antonio Foncubierta-Rodriguez** [1] **Florinda Feroce** [4]
Giosue Scognamiglio [4] **Anna Maria Anniciello** [4] **Jean-Philippe Thiran** [2] **Orcun Goksel** [3] **Maria Gabrani** [1]

Abstract

Explainability of machine learning (ML) techniques in digital pathology (DP) is of great significance to facilitate their wide adoption in clinics. Recently, graph techniques encoding relevant biological entities have been employed to represent and assess DP images. Such paradigm shift from pixel-wise to entity-wise analysis provides more control over concept representation. In this paper, we introduce a post-hoc explainer to derive compact per-instance explanations emphasizing diagnostically important entities in the graph. Although we focus our analyses to cells and cellular interactions in breast cancer subtyping, the proposed explainer is generic enough to be extended to other topological representations in DP. Qualitative and quantitative analyses demonstrate the efficacy of the explainer in generating comprehensive and compact explanations.

1. Introduction

Convolutional Neural Networks (CNNs), so far the most successful ML approach in image analysis, have been widely adopted to assess DP images to improve diagnosis and patient outcome. However, concept representations of CNNs remain unexplained in DP and thus hinder their adoption in typical workflows. Therefore, explainable ML technologies in DP have become of paramount interest to build trust and promote the employment of ML in clinical settings (Holzinger et al., 2017).

Typically CNNs process complex and large DP images in a patch-wise manner, followed by aggregating the patch-wise learning to address downstream DP tasks. Recently, several research works have been devoted to demystify the concept representations of CNNs in automated diagnosis. Patch-level explainable methods (Graziani et al., 2018; Hägele et al., 2019; Bruno et al., 2017; Mobadersany et al., 2017; Cruz-Roa et al., 2013; Xu et al., 2017) build patch-level *heatmaps*, where an importance score is computed per pixel to identify the regions of importance. For instance, Hägele et al. (2019) use layer-wise relevance propagation (Bach et al., 2015) to generate positive scores for pixels that are positively correlated with the class label and negative scores otherwise. Such approaches have several limitations. First, pixel-level heatmaps fail to capture the spatial organization and interactions of relevant biological entities. Second, the pixel-level analysis is completely detached from any biological reasoning that pathology guidelines recommend for decision making. Third, pixel-level explanation are common in the form of blurry heatmaps, which then do not allow to discriminate the relevance of nearby entities and their interactions.

Recently, graph techniques have been adopted to map DP patches to graph representations and process such graphs for pathology tasks (Gunduz et al., 2004; Zhou et al., 2019; Sharma et al., 2016; Gadiya et al., 2019; Wang et al., 2019; Pati et al., 2020). Graph representations embed biological entities and their interactions. To the best of our knowledge, explainability of *graph-based* approaches for DP has not been addressed yet. In this paper, a major step towards explainability in DP is presented based on two proposals: First, we advocate for shifting the analysis from a pixel-level representation to a relevant biological entity/relationship-oriented representation. The learning can then be regulated to specific entities and interactions, aligned with the biological and pathological knowledge. Second, we propose to adopt an instance-level post-hoc explainability method that extracts a relevant subset of entities and interactions from the input graph. We define this subset as the explanation of our original graph analysis. We hypothesize that the explanation will be deemed useful if and when the subset aligns with prior pathological knowledge. In this paper, we map DP images to cell-graphs (Gunduz et al., 2004), where cells and cellular interactions are represented as nodes and edges of the graph, and focus on the interpretability of cell-graphs towards cancer subtyping.

*Equal contribution [1] IBM Research Zürich, Zürich, Switzerland [2] Institute of Electrical Engineering, EPFL, Lausanne, Switzerland [3] Department of Information Technology and Electrical Engineering, ETH Zürich, Zürich, Switzerland [4] National Cancer Institute - IRCCS-Fondazione Pascale, Naples, Italy. Correspondence to: Guillaume Jaume <gja@zurich.ibm.com>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).

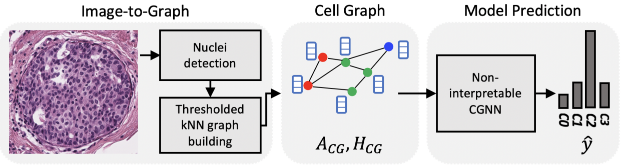


Figure 1: A RoI is transformed into a CG, and is processed by CGNN to predict the cancer subtype.

2. Methodology

In this section, we first present the extraction of *graph* representations from DP images, and further present the Graph Neural Network (GNN) framework for the processing of the representations. Second, we introduce the explainability module to acquire comprehensive explanations.

2.1. Cell-graph representation and learning

The DP images are transformed into cell-graph (CG) representations. Formally, we define a CG, $G_{\text{CG}} = (V, E, H)$ as an undirected graph composed of a vertices V and edges E. Each vertex is described by an embedding $h \in \mathbb{R}^d$, or equivalently expressed in its matrix form as $H \in \mathbb{R}^{|V| \times d}$. The graph topology is described by a symmetric adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$, where $A_{u,v} = 1$ if an edge exists between vertices u and v.

To build CG, we detect nuclei at 40× resolution using Hover-Net (Graham et al., 2019), a state-of-the-art nuclei segmentation algorithm pre-trained on MoNuSeg dataset (Kumar & et al., 2020). We extract 16 hand-crafted features incorporating shape, texture and color attributes to represent each nucleus as in (Zhou et al., 2019). We include centroid location normalised by the image size to spatially encode the nucleus. The detected nuclei and their 18-dimensional embeddings serve as the node and initial node embeddings of our CG. The CG topology assumes that spatially close cells encode biological interactions and consequently should form an edge. We use the k-Nearest Neighbors (kNN) algorithm, *i.e.*, for each node u, we build edges e_{uv} to the k closest vertices v. As isolated cells have weak cellular interaction with other cells, they ought to stay detached. Thus, we threshold the kNN graph by removing edges that are longer than a specified distance. We set $k = 5$ and the distance threshold to 50 pixels in our modeling.

For the downstream DP task, we determine the breast cancer subtypes of regions-of-interest (RoIs). For a dataset with N RoIs, we create $\mathcal{D} = \{G_{\text{CG},i}, l_i\}_{i=\{1,\ldots,N\}}$ consisting of N CGs and corresponding cancer stage labels l_i. A GNN (Defferrard et al., 2016; Kipf & Welling, 2017; Veličković et al., 2018; Xu et al., 2019), denoted as CGNN, is employed to build fixed-size graph embeddings from the CGs. These embeddings are fed to a Multi-Layer Perceptron (MLP) to predict the cancer stages. In particular, we use the Graph Isomorphism Network (GIN) (Xu et al., 2019), an instance of message passing neural network (Gilmer et al., 2017). A block diagram with the main steps is presented in Figure 1.

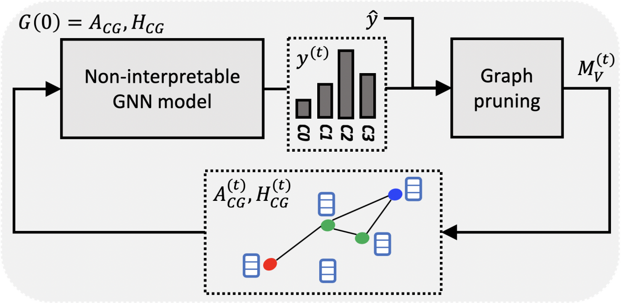


Figure 2: Overview of CGExplainer. The original CG is iteratively pruned until convergence of the optimization.

2.2. Cell-graph explainer

We propose a cell-graph explainer (CGExplainer) inspired by the GnnExplainer (Ying et al., 2019), a post-hoc interpretability method based on a graph pruning optimization. Considering the large number of cells in a RoI, we hypothetize that many of them will provide little information in the decision making, whereas others will be responsible for class specific patterns that would allow better understanding of the disease. Thus, we prune the redundant and uninformative graph components, and define the resulting sub-graph as the *explanation*.

Formally, let us consider a trained GNN model $\mathcal{M}$, and a sample $\{G_{\text{CG}}, l\}$ from $\mathcal{D}$ predicted as $\hat{y} = \mathcal{M}(G_{\text{CG}})$. We aim to find a sub-graph $G_s = (V_s, E_s, H_s) \subset G_{\text{CG}}$ such that the mutual information between the original prediction and the sub-graph is maximized, *i.e.*,

$$\max_{G_s} \text{MI}(\hat{Y}, G_s) = \mathcal{H}(\hat{Y}) - \mathcal{H}(\hat{Y}|G_{\text{CG}} = G_s) \quad (1)$$

which is equivalent to minimizing the conditional entropy:

$$\mathcal{H}(\hat{Y}|G_{\text{CG}} = G_s) = -\,\mathbb{E}_{\hat{Y}|G_s}[\log(P_{\mathcal{M}}(\hat{Y}|G_s))] \quad (2)$$

Intuitively, G_s maximizes the probability of $\hat{y}$. Direct optimization of Equation (2) is intractable due the combinatorial nature of graphs. Therefore, the GnnExplainer proposes to learn a mask that activates or deactivates parts of the graph. Considering the coherent pathological explainability of cells compared to cellular interactions, we focus on interpreting the *cells* in this work. Thus, we aim at learning a mask M_V at *node-level* that satisfies:

$$\min_{M_V} -\sum_{c=1}^{C} \mathbb{1}_{[y=c]} \log(P_{\mathcal{M}}(\hat{Y}|G_{\text{CG}}, \sigma(\text{diag}(M_V))H))) \quad (3)$$

where C denotes the number of classes, σ is the sigmoid activation, and $\mathrm{diag} : \mathbb{R}^{|V|} \rightarrow \mathbb{R}^{|V|\times|V|}$ is the diagonal matrix of the weight vector M_V. We intend the explanations to be as compact as possible, ideally with binarized weights, while providing the same prediction as the original graph. Heuristically, we enforce these constraints by minimizing:

$$\mathcal{L} = \mathcal{L}_{\mathrm{KD}}(\hat{y}, y^{(t)}) + \alpha_{M_V} \sum_{i}^{|V|} \sigma(M_{V_i}^{(t)}) + \alpha_{\mathcal{H}} \mathcal{H}^e(\sigma(M_V^{(t)})) \tag{4}$$

where, t is the optimization step. First term is the knowledge-distillation loss $\mathcal{L}_{\mathrm{KD}}$ between the new logits $y^{(t)}$ and the original prediction $\hat{y}$. Second term ensures the compactness of M_V. Third term binarizes M_V by minimizing its element-wise entropy $\mathcal{H}^e$. Following (Hinton et al., 2015), $\mathcal{L}_{\mathrm{KD}}$ is a combination of distillation and cross-entropy loss:

$$\mathcal{L}_{\mathrm{KD}} = \lambda \mathcal{L}_{\mathrm{CE}} + (1-\lambda)\mathcal{L}_{\mathrm{DIST}} \text{ where } \lambda = \frac{\mathcal{H}^e(y^{(t)})}{\mathcal{H}^e(\hat{y})} \tag{5}$$

As the element-wise entropy $\mathcal{H}^e(y^{(t)})$ increases, $\mathcal{L}_{\mathrm{CE}}$ gains importance and avoids a change in predicted label. M_V, produced by optimizing Equation (4), identifies important nodes with a weight factor. An overview of the explainer module is shown in Figure 2.

3. Experiments

3.1. Dataset

We evaluate CGEXPLAINER on BRACS dataset, an inhouse collection of BReAst Carcinoma Subtyping[1] images. The dataset consists of 2080 RoIs acquired from 106 H&E stained breast carcinoma whole-slide-images (WSIs). The RoIs are extracted at 40× magnification producing images of various sizes and appearances. The RoIs are annotated by the consensus of three pathologists as: normal (N), benign[2] (B), atypical[3] (A), ductal carcinoma in situ (D), and invasive (I) (a 5-class problem). We also study two simplified scenarios: (1) a 2-class problem: benign (N+B) and malignant (D+I) categories, and (2) a 3-class problem: benign (N+B), atypical (A), and malignant (D+I) categories. These scenarios allow us to study the relation between the task complexity and the generated explanations. Non-overlapping train, validation and test splits are created at WSI-level consisting of 1356, 365, and 359 RoIs respectively.

3.2. Implementation details

The experiments are conducted using PyTorch (Paszke et al., 2019) and the DGL library (Wang et al., 2019). The CGNN consists of three GIN layers with a hidden dimension of 32. Each GIN layer uses a 2-layer MLP with ReLU activation. The classifier consists of a 2-layer MLP with 64 hidden neurons that maps the hidden dimensions to the number of classes. The model is trained using the Adam optimizer with an initial learning rate of 10^{-3} and a weight decay of 5×10^{-4}. The batch size is set to 16.

The explanation module uses the trained CGNN. The mask M_V is learned by using the Adam optimizer with a learning rate of 0.01. The size constraint and the entropy constraint contribute to the loss by weighting factors $\alpha_{M_V} = 0.005$ and $\alpha_{\mathcal{H}} = 0.1$ respectively. The weights are adjusted such that the individual losses have comparable range. An early stopping mechanism is triggered, if G_s predicts a different label before reaching convergence. This ensures that the graph and its explanation always have the same prediction.

3.3. Quantitative and qualitative analyses

We conduct absolute and comparative analyses between CGEXPLAINER and random-explainer (RGEXPLAINER). RGEXPLAINER generates a random explanation from an original CG for a RoI by retaining equal number of nodes and edges as retained by CGEXPLAINER. We quantitatively and qualitatively evaluate the explainers under 2, 3 and 5-class scenarios, and assess them using surrogate metrics in absence of ground truth explanations. Table 1 presents the weighted F1-scores for CGNNs, the average node and edge reduction in CGEXPLAINER explanations, and cross-entropy (CE) loss of CGNN for processing the original CG, CGEXPLAINER-based CG and RGEXPLAINER-based CG. The cross-entropy is computed between the predicted logits and ground-truth labels of the RoIs.

The CGEXPLAINER removes a large percentage of nodes and edges to generate compact explanations across 2, 3 and 5-class scenarios while preserving the RoI predictions. The decrease in the percentage of node reduction with the increase in the number of classes indicates that with the increment of task complexity, the explainer exploits more nodes to extract valuable information. A similar pattern is observed for the edge reduction. Further, the reduction percentage decreases with the increase in the malignancy of the RoI. This indicates that the explainer discards abundantly available less relevant benign epithelial, stromal and lymphocytes, and retains relevant tumor and atypical nuclei. Combining the CG explanations in Figure 3 and the nuclei types annotation in Figure 4, we infer that the explanations retain relevant tumor epithelial nuclei for DCIS diagnosis. For 2-class scenario, the CG includes tumor nuclei in the central region of the gland. Few tumor nuclei are sufficient to differentiate (D) from (N+B). For 3-class scenario, the CG includes more tumor nuclei in the central region and the periphery of the gland and does not consider atypical

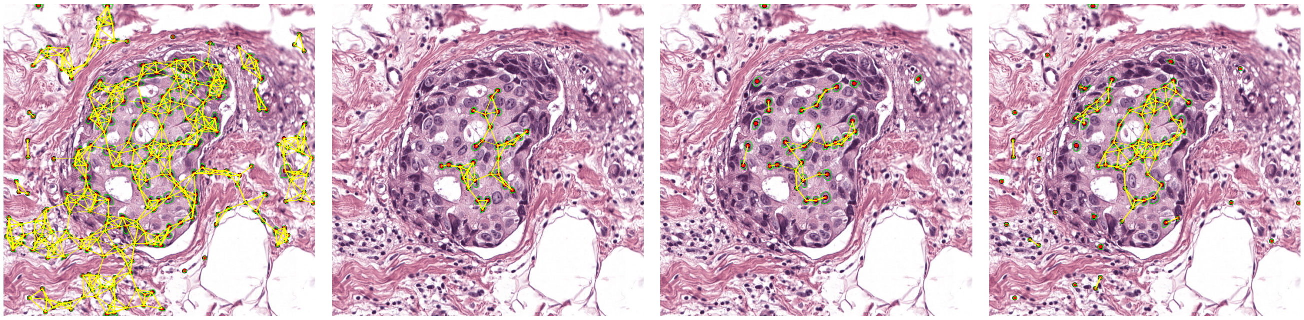
[1] currently pending approval for releasing the dataset

[2] includes benign and usual ductal hyperplasia

[3] includes flat epithelial atypia and atypical ductal hyperplasia

Metric/Scenario	**2-class scenario**			**3-class scenario**				**5-class scenario**					
	N+B	D+I	All	N+B	A	D+I	All	N	B	A	D	I	All
Weighted F1-score (↑)	0.97	0.97	0.97	0.95	0.35	0.80	0.77	0.56	0.74	0.37	0.62	0.77	0.61
Node reduction (%) (↑)	97.7	91.6	94.6	89.5	92.4	85.6	88.5	92.3	93.8	75.8	63.3	59.00	76.9
Edge reduction (%) (↑)	99.2	93.8	96.4	94.3	98.7	90.5	93.5	97.1	97.0	90.6	74.1	62.8	84.2
Original CE (↓)	0.21	0.21	0.21	0.45	2.05	0.38	0.72	2.65	0.59	2.22	0.72	0.48	1.21
Explanation CE (↓)	0.10	0.21	0.16	0.44	1.41	0.55	0.67	1.65	0.73	1.61	2.57	0.67	1.41
Random CE (↓)	0.02	3.14	1.61	1.00	0.38	1.75	1.20	0.62	0.93	1.52	11.4	2.85	3.55

Table 1: Quantitative results for CGNN, CGExplainer compactness, CGExplainer and RGExplainer performances.



(a) Original CG

(b) 2-class explanation CG

(c) 3-class explanation CG

(d) 5-class explanation CG

Figure 3: Qualitative comparison of original CG and CGExplainer CGs for 2, 3 and 5-class scenarios for a DCIS RoI.

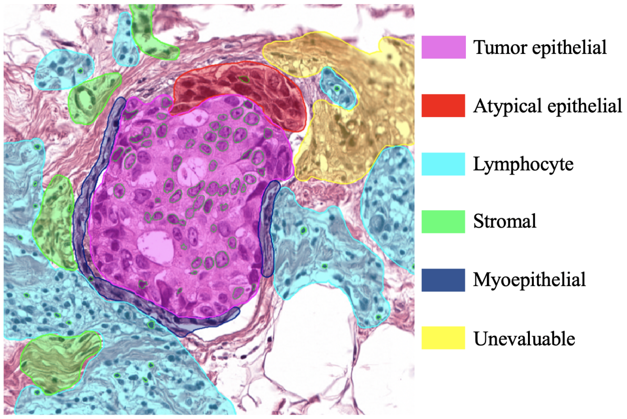


Figure 4: Nuclei types annotation. Overlaid segmentation masks of nuclei from 5-class explanation in green.

nuclei. This pattern differentiates (D) from (A). For 5-class scenario, the CG includes more tumor nuclei distributed within and around the gland, and some lymphocytes around the gland. The CG also includes more cellular interactions to identify a large cluster of tumor nuclei. Pathologically this behavior differentiates (D) from (I) which has small clusters of tumor nuclei scattered throughout the RoI. Additionally, the retained tumor nuclei and their interactions are consistent with increasing task complexity.

Further, we compare the class-wise logits for original, CGExplainer and RGExplainer CG via cross-entropy (CE). Table 1 presents the class-wise CE and average CE across *all* the classes. The CGExplainer-based CG and the original CG have comparable class-wise CE and average CE across all scenarios. We observe that in each scenario, the RGExplainer-based CG is biased towards one class. For instance, in the 2-class scenario, RGExplainer frequently predicts the class (N+B) leading to a per-class CE smaller than CGExplainer. However, on average across *all* the classes, the RGExplainer CE is consistently higher than the CGExplainer. This conveys that the RGExplainer removes relevant entities from CGs, thereby increasing the loss. These qualitative and quantitative analyses conclude that the CGExplainer generates meaningful and consistent explanations.

4. Conclusion

We believe that our work, though preliminary, is a step in the right direction towards better representations and interpretability in DP. We have herein focused on the methodological introduction and cell-level analyses. In future work, we plan to extend our approach to other biological entities and further to pathological assessment. Ultimately, our goal is to understand any information additional to an ML model prediction that one needs to provide to a user, to build trust and to facilitate adoption and deployment of such ML technologies in clinical scanarios.

Acknowledgement

We thank Nadia Brancati, Maria Frucci and Daniel Riccio, our colleagues from Institute for High Performance Computing and Networking - CNR, Italy, for their insights and expertise that greatly assisted the research. We are also immensely grateful to Maurizio Do Bonito and Gerardo Botti from National Cancer Institute - IRCCS-Fondazione Pascale, Italy, and Giuseppe De Pietro from Institute for High Performance Computing and Networking - CNR, Italy for their support and collaborative research with IBM Research Zurich, Switzerland.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), jul 2015. ISSN 19326203. doi: 10.1371/journal.pone.0130140.

Bruno, K., Andrea, M. O., Allen, P. M., Catherine, M. N., Matthew, A. S., Lorenzo, T., Arief, A. S., and Saeed, H. Looking under the hood Deep neural network visualization to interpret whole slide Image analysis outcomes for colorectal polyps. In *CVPR-W*, 2017.

Cruz-Roa, A., Ovalle, A. A., John, E., Madabhushi, A., and González Osorio, F. A. LNCS 8150 - A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection. In *MICCAI*, 2013.

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *NIPS*, number Nips, 2016. ISBN 978-1-5108-3881-9. URL `http://arxiv.org/abs/1606.09375`.

Gadiya, S., Anand, D., and Sethi, A. Histographs: Graphs in histopathology. *arXiv preprint arXiv:1908.05020*, 2019.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *International Conference on Machine Learning, ICML*, volume 70, pp. 1263–1272, 2017.

Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., and Rajpoot, N. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

Graziani, M., Andrearczyk, V., and Müller, H. Regression Concept Vectors for Bidirectional Explanations in Histopathology. In *MICCAI-iMIMIC*, apr 2018. URL `http://arxiv.org/abs/1904.04520`.

Gunduz, C., Yener, B., and Gultekin, S. H. The cell graphs of cancer. *Bioinformatics*, 20(suppl_1):i145–i151, 2004.

Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.-R., and Binder, A. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. In *MICCAI-iMIMIC*, aug 2019. URL `http://arxiv.org/abs/1908.06943`.

Hinton, G., Vinyals, O., and Dean, J. Distilling the Knowledge in a Neural Network. In *NeurIPS*, mar 2015. URL `http://arxiv.org/abs/1503.02531`.

Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., and Zatloukal, K. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. dec 2017. URL `http://arxiv.org/abs/1712.06657`.

Kipf, T. N. and Welling, M. Semi supervised classification with graph convolutional networks. In *ICLR*, pp. 1–14, 2017.

Kumar, N. and et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5):1380–1391, 2020.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D., Barnholtz-Sloan, J., Velazquez Vega, J. E., Brat, D., and Cooper, L. A. Predicting cancer outcomes from histology and genomics using convolutional networks. *PNAS*, pp. 198010, 2017. doi: 10.1101/198010.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., and Others. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Pati, P., Jaume, G., Alisha Fernandes, L., Foncubierta, A., Feroce, F., Anniciello, A. M., Scognamiglio, G., Brancati, N., Riccio, D., Do Bonito, M., De Pietro, G., Botti, G., Goksel, O., Thiran, J.-P., Frucci, M., and Gabrani, M. HACT-Net: A Hierarchical Cell-to-Tissue Graph Neural Network for Histopathological Image Classification. 2020.

Sharma, H., Zerbe, N., Heim, D., Wienert, S., Lohmann, S., Hellwich, O., and Hufnagl, P. Cell nuclei attributed relational graphs for efficient representation and classification of gastric cancer in digital histopathology. In *SPIE Medical Imaging 2016: Digital Pathology*, volume 9791, pp. 97910X, 2016.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph Attention Networks. *ICLR*, 2018. URL `http://arxiv.org/abs/1710.10903`.

Wang, J., Chen, R. J., Lu, M. Y., Baras, A., and Mahmood, F. Weakly supervised prostate tma classification via graph convolutional networks. *arXiv preprint arXiv:1910.13328*, 2019.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations, ICLR*, 2019.

Xu, Y., Jia, Z., Wang, L. B., Ai, Y., Zhang, F., Lai, M., and Chang, E. I. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*, 18(1), may 2017. ISSN 14712105. doi: 10.1186/s12859-017-1685-x.

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. In *NeurIPS*, 2019. URL `http://arxiv.org/abs/1903.03894`.

Zhou, Y., Graham, S., Koohbanani, N. A., Shaban, M., Heng, P.-A., and Rajpoot, N. CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images. In *ICCV-W*, 2019. URL `http://arxiv.org/abs/1909.01068`.