Enhancing Whole Slide Image Classification with Discriminative and Contrastive Learning

Peixian Liang, Hao Zheng, Hongming Li, Yuxin Gong, Yong Fan

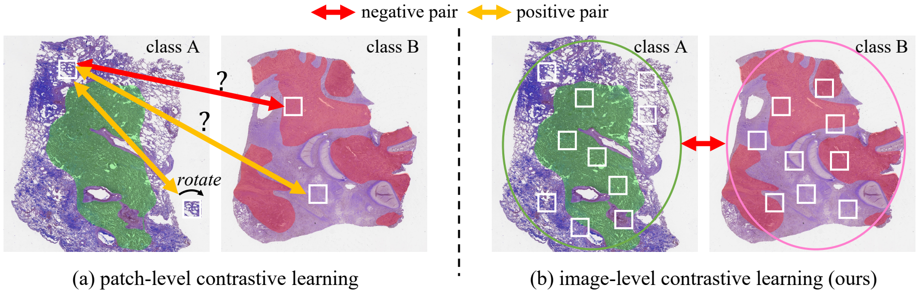
University of Pennsylvania
{peixian.liang,hao.zheng,hongming.li,yuxin.gong,yong.fan}@pennmedicine.upenn.edu

Abstract. Whole slide image (WSI) classification plays a crucial role in digital pathology data analysis. However, the immense size of WSIs and the absence of fine-grained sub-region labels, such as patches, pose significant challenges for accurate WSI classification. Typical classification-driven deep learning methods often struggle to generate compact image representations, which can compromise the robustness of WSI classification. In this study, we address this challenge by incorporating both discriminative and contrastive learning techniques for WSI classification. Different from the extant contrastive learning methods for WSI classification that primarily assign pseudo labels to patches based on the WSI-level labels, our approach takes a different route to directly focus on constructing positive and negative samples at the WSI-level. Specifically, we select a subset of representative and informative patches to represent WSIs and create positive and negative samples at the WSI-level, allowing us to better capture WSI-level information and increase the likelihood of effectively learning informative features. Experimental results on two datasets and ablation studies have demonstrated that our method significantly improved the WSI classification performance compared to state-of-the-art deep learning methods and enabled learning of informative features that promoted robustness of the WSI classification.

1 Introduction

Digital scans of pathology tissue slides, often referred to as whole slide images (WSIs), provide rich information, such as tumor microenvironments, for cancer diagnosis and treatment planning [1, 18]. While WSI classification plays an important role in addressing cancer diagnosis, it presents a significant challenge due to the gigapixel size of WSIs and the absence of pixel-level annotations.

Deep learning methods for the WSI classification task typically divide the huge WSIs into image patches and integrate the image patches for classification at the WSI-level based on features extracted from the image patches [24, 4, 3, 13]. Promising WSI classification performance has been achieved by deep learning methods with innovative graph and Transformer-based architectures that facilitate effective feature learning and patch integration for the WSI classification [5, 3, 21, 12]. Despite their promising classification performance, these classifier-driven methods face challenges in attaining compact image representations to enhance the robustness of classification accuracy in that these methods



(a) patch-level contrastive learning

(b) image-level contrastive learning (ours)

Fig. 1. A comparison of image-level contrastive learning with patch-level contrastive learning, given two WSIs from two different cancer classes, A and B, respectively. (a) for the patch-level contrastive learning, positive and negative pairs of image patches are needed. However, the absence of patch-level label information introduces potential noise in both positive and negative pairs, while self-supervised contrast learning cannot utilize class label information. (b) for the image-level constrastive learning, positive and negative samples are defined based on the class label information of WSIs, and such information is propagated to the image patches for enhancing feature learning with our proposed method. By treating a set of patches as the basic unit, it allows to learn a more comprehensive representation of WSIs and increases the likelihood of capturing cancerous regions. Cancer areas are depicted by the green and red colors in the corresponding images.

employ discriminative information alone to learn features and construct classification models, ignoring intra-class and inter-class feature variabilities [19, 30].

We aim to address this challenge and obtain compact and informative image representations for accurate WSI classification through join discriminative and contrastive learning. Contrastive learning is an effective method to learn compact feature representations by minimizing feature distances between positive samples while maximizing distances across negative samples. Existing contrastive learning methods for WSI classification can be broadly categorized into two types: self-supervised learning and weakly supervised learning. In the self-supervised methods, patches, along with their augmented or semantically similar counterparts, are regarded as positive samples while semantically dissimilar patches are considered as negative samples [11, 27, 25]. Despite the rich semantic information obtained by these methods, there exists a limitation in exploring pathology-related discriminative information since the positive and negative samples are not tied to WSI class information. In the weakly supervised learning methods, image-level labels are utilized to assign pseudo class labels to patches, forming the basis for the construction of positive and negative patch pairs [26, 22]. However, transferring the WSI-level label information to image patches may introduce class label noise and yield degraded image representations. Taking the

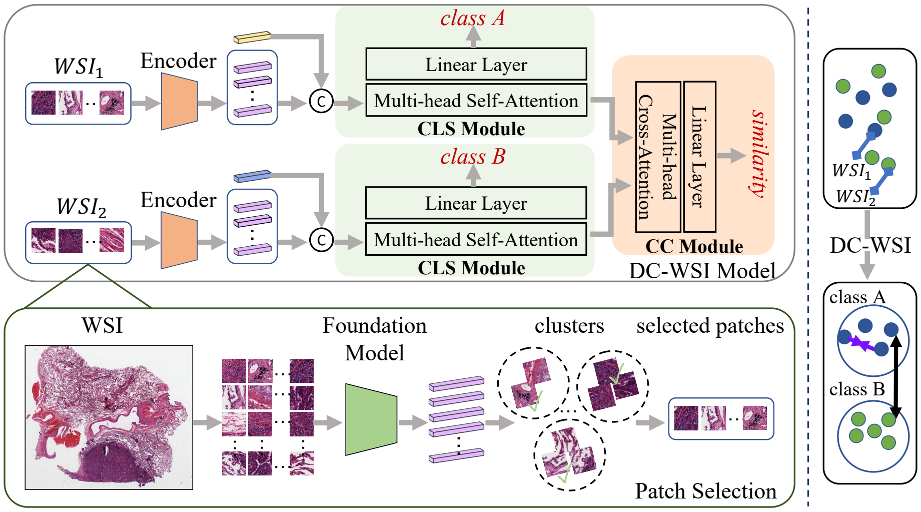


Fig. 2. An overview of our proposed DC-WSI method. Two WSIs are sampled from the training set for demonstration. Given each WSI representation (i.e., a set of selected patches), an encoder is applied to extract patch features. The classification (CLS) module aggregates intra-image features to predict class labels, and the contrastive learning (CC) module facilitates effective learning of informative features that maximize intra-class similarity and minimize inter-class similarity.

abnormal WSI images for example: both normal and abnormal patches may co-exist. Assigning abnormal pseudo labels to normal patches of the abnormal WSI can introduce extraneous noise in subsequent contrastive learning.

To overcome limitations of the extant methods, we introduce a novel framework, **DC-WSI**: **D**iscriminative and **C**ontrastive learning framework for **W**hole **S**lide **I**mage classification. Our approach employs both discriminative and contrastive learning to obtain compact and robust image representation at both the patch- and the WSI-levels for accurate WSI classification in an end-to-end multi-task WSI classification framework. Specifically, our method employs attention mechanisms to aggregate patch features for making image classification predictions. Our contrastive learning uses a set of patches to represent a WSI, with the WSI class label (discriminative) information propagated to the image patches through cross-attention, which facilitates effective learning of more robust representations of the WSIs and improves the likelihood of detecting the abnormal patches of WSIs (see Fig. 1). Such a contrastive learning strategy allows to aggregate patch features of WSIs through cross-attention for characterizing similarities between WSI images and encouraging feature learning to maximize intra-class similarity and minimize inter-class similarity.

Our contributions are three-folds: ***1)*** We propose a new discriminative and contrastive learning framework to learn compact and robustness image representations for accurate WSI classification; ***2)*** We propose a new WSI-level contrastive learning method with a set of patches to refrain from using patch-level pseudo labels and thus mitigate the label noise in the learning process; and *3)* We design a deep learning model with joint discriminative and contrastive learning to improve the WSI classification performance.

2 Method

2.1 Problem Definition

Given a set of WSIs $X = \{X_n\}_{n=1}^{N}$, which is split into two parts: training set $X_{train} = \{X_i\}_{i=1}^{M}$ and testing set $X_{test} = \{X_j\}_{j=M+1}^{N}$. Each training WSI $X_i \in X_{train}$ has its binary image class label $Y_i \in \{0, 1\}$ representing normal/abnormal or different disease types. Our goal is to predict class label $Y_j \in \{0, 1\}$ for each test image $X_j \in X_{test}$.

2.2 Method Overview

Our method is schematically illustrated in Fig. 2, consisting of two parts: *Patch Selection* and *DC-WSI Model*. (1) *Patch Selection*: Selecting a subset of informative and representative patches to represent each of the WSIs to facilitate computationally efficient WSI classification. Specifically, given a WSI X_i, we divide it into m non-overlapping patches (m can vary for different WSIs). A foundation model is then applied to encode each patch into a fixed dimensional vector to capture the semantic information of the patch. All vectors within a WSI are then input into a clustering method to group the patches into k clusters. Subsequently, q patches are randomly selected from each cluster, forming a set $P_i = \{p_{i,1}, p_{i,2}, ..., p_{i,b}\}$, where b is the number of selected patches. P_i is used for further training. (2) *DC-WSI Model*: We construct an end-to-end dual discriminative and contrastive learning model to predict class labels of WSIs. For the contrastive learning, we sample a pair of WSIs, $X_1, X_2 \in X_{train}$ from the training set. If they belong to the same class, their corresponding patch set P_1 and P_2 are considered as a positive pair; otherwise, it is a negative pair. An encoder is then applied to each patch $p_{i,j}, i \in \{1, 2\}$ to produce a fixed-dimensional vector $f_{i,j}$ that captures disease-aware information. After encoding, we transform patch set P_i into the encoding space $F_i = \{f_{i,j}\}_{j=1}^{b}$. Subsequently, a self-attention module aggregates features F_i to make a classification prediction for X_i. Additionally, a contrastive learning module using cross-attention layers to characterize similarity between X_1 and X_2 with a similarity score $sim_{1,2}$, based on their patch features F_1 and F_2. The contrastive learning loss is designed to encourage positive samples to be similar to each other while pushing negative samples apart.

2.3 Representative Patch Selection

We apply SAM [10] as the feature extractor to obtain patch features for subsequent clustering. SAM is a robust foundation model capable of extracting dataset-agnostic semantic information. Specifically, we employ the SAM encoder to transform patches into fixed-size one-dimensional vectors. Once all patch features within a WSI are obtained, we apply the K-means clustering method [16] to categorize the corresponding patches into k clusters. For each cluster, we randomly sample q patches. The collection of selected patches form a patch set $P_i = \{p_{i,1}, p_{i,2}, ..., p_{i,b}\}$, where $b <= m$. P_i denotes a WSI X_i to be used for feature learning and WSI classification.

2.4 DC-WSI Model

Encoder We employ ResNet18 [7] as an encoder backbone, excluding its last three layers. Given a WSI with informative and representative patches $P_i = \{p_{i,j}\}_{j=1}^{b} \in R^{b \times h \times w}$, where $h \times w$ is the patch size, it is used as an input to the encoder for learning patch features: $F_i = \{f_{i,j}\}_{j=1}^{b} \in R^{b \times h' \times w'}$, where $h' \times w'$ represents feature map size. These patch features are then flattened to produce $F_i = \{f_{i,j}\}_{j=1}^{b} \in R^{b \times d}$, where $d = h' \times w'$. F_i is to be optimized through training by minimizing a WSI classification loss and a contrastive learning loss in an end-to-end training process.

Classification Module The classification (CLS) module aims to predict a class label of X_i based on its patch features F_i. Specifically, following the setting of ViT [6], we add a learnable class token $C \in R^{1 \times d}$ into the patch features F_i to learn a set of features $F_i' = [F_i; C] \in R^{(b+1) \times d}$. Then, F_i' is fed into a Multi-head Self-Attention module. Specifically, F_i' goes through a *multi-head attention* [23] layer, which yields query, key, and value vectors: $Q_i = F_i' \times W^Q, K_i = F_i' \times W^K, V_i = F_i' \times W^V$, where $W^Q, W^K, W^V \in R^{d \times d}$ are parameter matrices. Finally, for each head j an attention output is computed as:

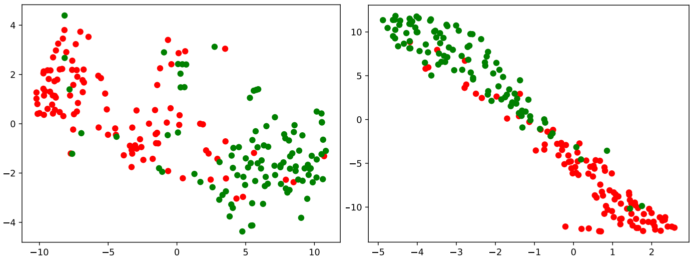
$$A_i^j = softmax(\frac{Q_i^j (K_i^j)^T}{\sqrt{d}}) \times V_i^j \in R^{(b+1) \times d/h},$$

where $Q_i = [Q_i^1, ..., Q_i^h]$, $K = [K_i^1, ..., K_i^h]$, $V = [V_i^1, ..., V_i^h]$, Q_i^j, K_i^j, $V_i^j \in R^{(b+1) \times d/h}$, $i = 1, ..., h$, and h is the number of heads.

The attention outputs from all heads are concatenated to form a feature set $A_i = [A_i^1, ..., A_i^h]$. The class token $C_i' \in R^{1 \times d}$ is taken from A_i and passed through a linear layer to generate a prediction of the class probability, denoted as $q_i \in R$. The prediction is supervised by the image label Y_i, and the classification loss L_{cls} is computed as a binary cross-entropy loss:

$$L_{cls}(q_i, Y_i) = -\sum_i (Y_i log q_i + (1 - Y_i) log(1 - q_i)).$$

Contrastive Learning Module The Contrastive Learning (CL) module operates on pairs of images X_i and X_j with corresponding class labels of Y_i and Y_j respectively, sampled from the training set X_{train}. The objective is to optimize



(a) L_{cls} (b) L_{cls} + L_{cc}

Fig. 3. t-SNE visualization of image features (class tokens) of WSIs learned by (a) a classification model (L_{cls}) and (b) a classification+contrastive model ($L_{cls} + L_{cc}$) on TCGA-Lung test set. Different colors represent samples in different classes.

the image features by maximize intra-class similarity and minimize inter-class similarity. Positive pairs contain images with the same class label, i.e., $Y_i = Y_j$, while negative pairs contain images from different classes.

Specifically, for a pair of images with features of A_i and A_j, a *multi-head* attention is used to obtain cross-attention between them for computing their similarity, $sim_{i,j} \in [0, 1]$, between their class tokens C_i and C_j with a linear layer.

During each training iteration, Z pairs of images are sampled, with $Z/2$ are negative samples, and $Z/2$ are positive samples. The similarity scores of all positive pairs are added to derive SIM_{pos}, while the similarity scores of all negative pairs are added to derive SIM_{neg}. The contrastive learning loss L_{cc} is calculated using the maximum-margin classification loss:

$$L_{cc}(\sum SIM_{pos}, \sum SIM_{neg}) = max(0, (\sum SIM_{neg} - \sum SIM_{pos}) + margin).$$

The final loss is the sum of the classification loss L_{cls} and the contrastive learning loss L_{cc}. The model is trained end-to-end. During inference, only the classification branch is utilized to predict class labels for input WSIs.

3 Experiments

We conduct experiments on two datasets: TCGA-Lung and TCGA-ESCA to demonstrate the effectiveness of the proposed WSI-CL method. Additionally, we perform ablation studies to demonstrate the effectiveness of key components in our WSI-CL method.

Datasets *1)* TCGA-Lung is a public dataset from National Cancer Institute Data Portal [2]. It includes two types of lung cancer, i.e., Lung Squamous Cell Carcinoma (TCGA-LUSC) and Lung Adenocarcinoma (TCGA-LUAD). A total of 1042 diagnostic WSIs were collected and randomly divided into training and testing sets with a ratio of 0.8:0.2. The training set contained 409 TCGA-LUSC

Table 1. WSI classification comparison results on TCGA-Lung and TCGA-ESCA datasets. RS denotes random sampling, PS denotes proposed patch selection. The **bold** score represents the best performance on the corresponding dataset.

	TCGA-Lung		TCGA-ESCA	
	Accuracy (%) ↑	AUC (%) ↑	Accuracy (%) ↑	AUC (%) ↑
ABMIL [8]	0.785	0.866	0.750	0.846
TransMIL [20]	0.823	0.905	0.781	0.922
GTP [31]	0.876	**0.956**	0.875	0.949
CLAM [14]	0.876	0.952	0.875	0.941
DC-WSI (RS+L_{cls})	0.828	0.905	0.875	0.933
DC-WSI (PS+L_{cls})	0.856	0.916	0.906	0.938
DC-WSI (PS+L_{cls} + L_{cc})	**0.880**	0.940	**0.938**	**0.965**

and 424 TCGA-LUAD, and the test set contained 103 TCGA-LUSC and 106 TCGA-LUAD. *2)* TCGA-ESCA is a dataset from National Cancer Institute Data Portal [2]. A total of 149 diagnostic slides are collected which contains two types of Oesophageal Carcinoma, i.e., Squamous cell carcinoma (SCC) and Adenocarcinoma (AD). We randomly split it into training and testing sets with a ratio 0.8:0.2. The training set contained 68 SCC and 49 AD, and the testing set contained 16 SCC and 16 AD.

Evaluation Metrics We used the standard WSI classification evaluation metrics, including *Accuracy* [28] and area under the curve (AUC) score [29]. Specifically, $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, where TP=True positive, TN=True negative, FP=False positive, FN=False negative.

Implementation Details In our experimental setup, WSIs were partitioned into patches of size 224x224 at 10X magnification. For the patch selection, we used following parameters: $k = 8$ and $q = 50$. In the contrastive learning module, we configured $Z = 6$. The learning rate was set to 0.0002, and the optimization was performed using the Adam optimizer [9]. The model was implemented using PyTorch [17].

Comparison Methods We compared our method with an array of state-of-the-art (SOTA) WSI classification methods. Particularly, ABMIL [8] is a multi instance learning (MIL) framework to aggregate instance features through attention for final bag-level prediction, TransMIL [20] is a Transformer based WSI classification framework to aggregate patch features by attention mechanism, GTP [31] is a Transformer-Graph based WSI classification framework with a patch-level contrastive learning to learn patch features, and CLAM [14] is a clustering-constrained attention MIL approach with a patch clustering loss to impose constraints and refine patch features in the WSI classification process.

3.1 WSI Classification Results and ablation studies

Table 1 shows WSI classification comparison results on TCGA-Lung and TCGA-ESCA datasets. Firstly, our method (*DC-WSI (PS+*L_{cls} + L_{cc}*)*) obtained the overall best WSI classification performance among all methods under comparison. Particularly, our method achieved substantial improvement on the TCGA-ESCA dataset, with a 6.30% increase in *Accuracy* and a 1.6% increase in AUC compared to the second-best method. On the TCGA-Lung dataset, our approach achieved a 0.4% increase in *Accuracy* over the second-best method. These results demonstrated the effectiveness of our method in extracting discriminative features and aggregating them for accurate WSI classification predictions.

Secondly, ablation studies demonstrated the effectiveness of key components of our method. *DC-WSI (RS+*L_{cls}*)* denotes the ablation study of representative patch selection in Section 2.3. Instead of using our proposed patch selection strategy, *DC-WSI (RS+*L_{cls}*)* randomly selected the same amount of patches with representative patch selection for further training. The results in Table 1 showed that our patch selection strategy *DC-WSI (PS+*L_{cls}*)* outperformed *DC-WSI (RS+*L_{cls}*)*, indicating that our patch selection strategy can capture more informative patches, leading to more accurate WSI classification.

DC-WSI (PS+L_{cls}*)* denotes the contrastive learning ablation study. Instead of using both discriminative and contrastive learning modules, *DC-WSI (PS+*L_{cls}*)* only used the classification module with L_{cls} loss. Table 1 shows results. The performance degradation of *DC-WSI (PS+*L_{cls}*)* compared to *DC-WSI (PS+*$L_{cls} + L_{cc}$*)* demonstrated the effectiveness of the contrastive learning component in improving the classification performance.

Fig. 3 shows a t-SNE [15] visualization comparison of image representations (i.e., class tokens C_i) obtained from *DC-WSI (PS+*L_{cls}*)* and *DC-WSI (PS+*$L_{cls} + L_{cc}$*)* models respectively, further demonstrating that the contrastive learning module can help learn more informative features that maximized the intra-class similarity and minimized the inter-class similarity, compared with the classification model with the discriminative learning alone.

4 Conclusion

We develop a new discriminative and contrastive learning framework for WSI classification. Experimental results on two WSIs datasets and ablation studies have demonstrated that the proposed method can learn discriminative features that improved WSI classification performance, maximized intra-class similarity, and minimized inter-class similarity. Specifically, our method selects a subset of informative and representative patches as the basic unit of WSIs, while positive and negative samples are directly constructed at the WSI-level for the contrastive learning. Compared to the extant patch-level-based contrastive learning for the WSI classification, utilization of a set of patches as a basic unit not only facilitates effective learning of robust features from the WSIs but also improves classification performance. Our method can be further improved by incorporating the patch selection in the end-to-end learning, though our current strategy offers the flexibility to use different clustering algorithms to select representative patches.

References

1. Barisoni, L., Lafata, K.J., Hewitt, S.M., Madabhushi, A., Balis, U.G.: Digital pathology and computational image analysis in nephropathology. Nature Reviews Nephrology **16**(11), 669–685 (2020)
2. Cancer Genome Atlas Research Network, J., et al.: The cancer genome atlas pan-cancer analysis project. Nat. Genet **45**(10), 1113–1120 (2013)
3. Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: 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 (2023)
4. Chikontwe, P., Nam, S.J., Go, H., Kim, M., Sung, H.J., Park, S.H.: Feature re-calibration based multiple instance learning for whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 420–430. Springer (2022)
5. Ding, S., Wang, J., Li, J., Shi, J.: Multi-scale prototypical transformer for whole slide image classification. In: International conference on medical image computing and computer-assisted intervention. pp. 602–611. Springer (2023)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
11. Li, B., Li, Y., Eliceiri, K.W.: 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 (2021)
12. Li, Y., Shen, Y., Zhang, J., Song, S., Li, Z., Ke, J., Shen, D.: A hierarchical graph v-net with semi-supervised pre-training for histological image based breast cancer classification. IEEE Transactions on Medical Imaging (2023)
13. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023)
14. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
16. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967)

17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
18. Robertson, S., Azizpour, H., Smith, K., Hartman, J.: Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. Translational Research **194**, 19–35 (2018)
19. Roth, K., Brattoli, B., Ommer, B.: Mic: Mining interclass characteristics for improved metric learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8000–8009 (2019)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
21. Shi, J., Tang, L., Li, Y., Zhang, X., Gao, Z., Zheng, Y., Wang, C., Gong, T., Li, C.: A structure-aware hierarchical graph-based multiple instance learning framework for pt staging in histopathological image. IEEE Transactions on Medical Imaging (2023)
22. Tan, J.W., Jeong, W.K.: Histopathology image classification using deep manifold contrastive learning. arXiv preprint arXiv:2306.14459 (2023)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Wang, S., Yang, D.M., Rong, R., Zhan, X., Fujimoto, J., Liu, H., Minna, J., Wistuba, I.I., Xie, Y., Xiao, G.: Artificial intelligence in lung cancer pathology image analysis. Cancers **11**(11), 1673 (2019)
25. Wang, X., Du, Y., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Retccl: clustering-guided contrastive learning for whole-slide image retrieval. Medical image analysis **83**, 102645 (2023)
26. Wang, X., Xiang, J., Zhang, J., Yang, S., Yang, Z., Wang, M.H., Zhang, J., Yang, W., Huang, J., Han, X.: Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. Advances in neural information processing systems **35**, 18009–18021 (2022)
27. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
28. Wikipedia: https://en.wikipedia.org/wiki/Accuracy_and_precision
29. Wikipedia: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve
30. Yang, H.M., Zhang, X.Y., Yin, F., Liu, C.L.: Robust classification with convolutional prototype learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3474–3482 (2018)
31. Zheng, Y., Gindra, R.H., Green, E.J., Burks, E.J., Betke, M., Beane, J.E., Kolachalama, V.B.: A graph-transformer for whole slide image classification. IEEE transactions on medical imaging **41**(11), 3003–3015 (2022)