Cross-Scale Attention Guided Multi-instance Learning for Crohn's Disease Diagnosis with Pathological Images

Ruining Deng[1], Can Cui[1], Lucas W. Remedios[1], Shunxing Bao[1], R. Michael Womick[2], Sophie Chiron[3], Jia Li[3], Joseph T. Roland[3], Ken S. Lau[1], Qi Liu[3], Keith T. Wilson[3,4], Yaohong Wang[3], Lori A. Coburn[3,4], Bennett A. Landman[1], and Yuankai Huo[1(✉)]

[1] Vanderbilt University, Nashville, TN 37215, USA
yuankai.huo@vanderbilt.edu
[2] The University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA
[3] Vanderbilt University Medical Center, Nashville, TN 37232, USA
[4] Veterans Affairs Tennessee Valley Healthcare System, Nashville, TN 37212, USA

Abstract. Multi-instance learning (MIL) is widely used in the computer-aided interpretation of pathological Whole Slide Images (WSIs) to solve the lack of pixel-wise or patch-wise annotations. Often, this approach directly applies "natural image driven" MIL algorithms which overlook the multi-scale (i.e. pyramidal) nature of WSIs. Off-the-shelf MIL algorithms are typically deployed on a single-scale of WSIs (e.g., 20× magnification), while human pathologists usually aggregate the global and local patterns in a multi-scale manner (e.g., by zooming in and out between different magnifications). In this study, we propose a novel cross-scale attention mechanism to explicitly aggregate inter-scale interactions into a single MIL network for Crohn's Disease (CD), which is a form of inflammatory bowel disease. The contribution of this paper is two-fold: (1) a cross-scale attention mechanism is proposed to aggregate features from different resolutions with multi-scale interaction; and (2) differential multi-scale attention visualizations are generated to localize explainable lesion patterns. By training ∼250,000 H&E-stained Ascending Colon (AC) patches from 20 CD patient and 30 healthy control samples at different scales, our approach achieved a superior Area under the Curve (AUC) score of 0.8924 compared with baseline models. The official implementation is publicly available at https://github.com/hrlblab/CS-MIL.

Keywords: Multi-instance Learning · Multi-scale · Attention mechanism · Pathology

1 Introduction

Digital pathology is relied upon heavily by clinicians to accurately diagnose Crohn's Disease (CD) [14,32]. Pathologists carefully examine biopsies at multiple scales through microscopes to examine morphological patterns [6], which is a laborious task.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
X. Li et al. (Eds.): MMMI 2022, LNCS 13594, pp. 24–33, 2022.
https://doi.org/10.1007/978-3-031-18814-5_3

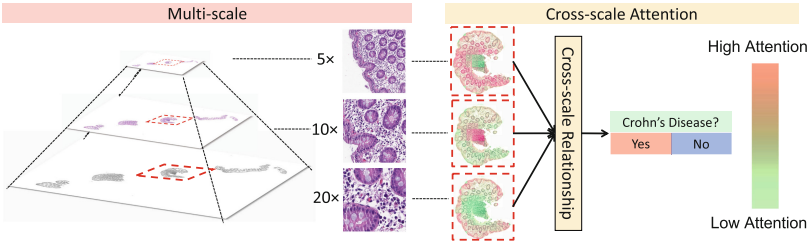

Fig. 1. Multi-scale awareness. Human pathologists typically aggregate the global and local patterns in a multi-scale manner. However, previous work failed to be aware of cross-scale relationship at different resolutions. Our method demonstrates the importance-of-regions with cross-scale attention maps, and aggregate the multi-scale patterns with differential attention scores for CD diagnosis.

With the rapid development of whole slide imaging (WSI) and deep learning methods, computer-assisted CD clinical prediction and exploration [9,18,19,27] are increasingly promising endeavors. However, annotating images pixel- or patch-wise is computationally expensive for a standard supervised learning system [11,16,23,24]. To achieve accurate diagnoses from weakly annotated images (e.g., patient-wise diagnosis), multi-instance Learning (MIL) – a widely used weakly supervised learning paradigm – has been applied to digital pathology [7,21,22,26,29]. For example, DeepAttnMISL [31] clustered image patches into different "bags" to model and aggregate diverse local features for patient-level diagnosis.

However, most prior efforts, especially the "natural image driven" MIL algorithms, ignore the multi-scale (i.e., pyramidal) nature of WSIs. For example, a WSI consists of a hierarchical scales of images (from 40× to 5×), which allows pathologists to examine both local [2] and global [1] morphological features [5,13,28]. More recent efforts have mimicked such human pathological assessments by using multi-scale images in a WSI [15,20]. These methods typically perform independent feature extraction at each scale and then perform a "late fusion". In this study, we consider the feasibility of examining the interaction between different scales at an earlier stage through an attention-based "early fusion" paradigm.

In this paper, we propose the addition of a novel cross-scale attention mechanism in an attention-guided MIL scheme to explicitly model inter-scale interactions during feature extraction (Fig. 1). In summary, the proposed method not only utilizes the morphological features at different scales (with different fields of view), but also learns their inter-scale interactions as a "early fusion" learning paradigm. Through empirical validation, our approach achieved the higher Area under the Curve (AUC) scores, Average Precision (AP) scores, and classification accuracy. The contribution of this paper is two-fold:

- A novel cross-scale attention mechanism is proposed to integrate the multi-scale information and the inter-scale relationships.
- Differential cross-scale attention visualizations are generated for lesion pattern guidance and exploration.

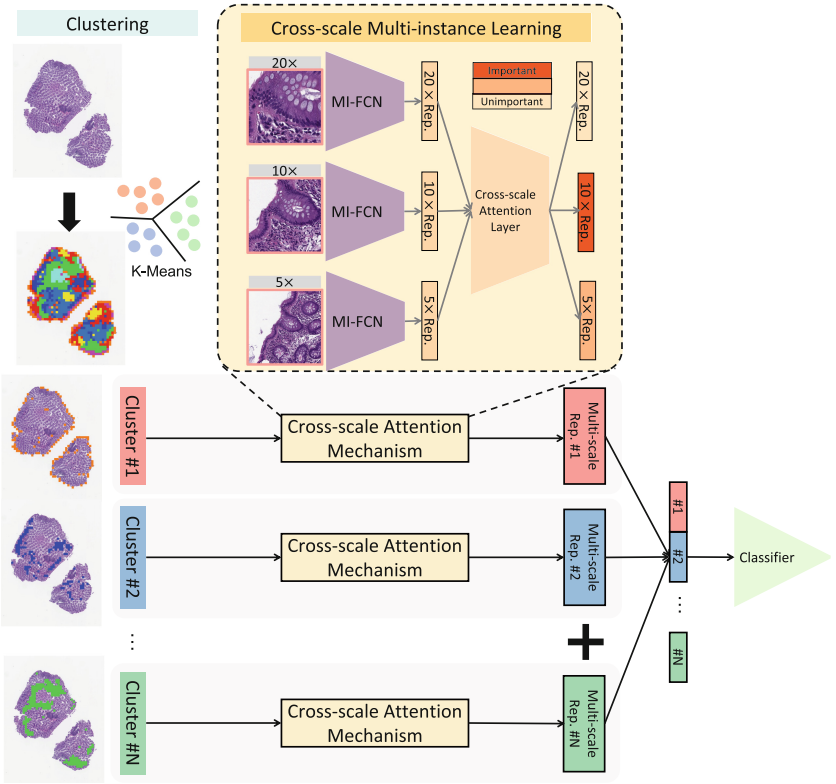

Fig. 2. Cross-scale Attention Guided Multi-instance Learning Pipeline. This figure demonstrates the pipeline of the proposed method. The local feature-based clustering was deployed on each WSI to distribute the phenotype patterns in each MIL bag. The cross-scale attention mechanism is deployed in each cluster of MIL branch to combine the multi-scale features with differential attention scores. Multi-scale representations from different clusters were concatenated for CD classification.

2 Methods

The overall pipeline of the proposed CS-MIL is presented in Fig. 2. Patches at each location (same center coordinates) at different scales are jointly tiled from WSIs. Patch-wise phenotype features are extracted from a self-supervised model. Then, local feature-based clustering is deployed on each WSI to distribute the phenotype patterns in each MIL bag. Cross-scale attention-guided MIL is proposed to aggregate features in multi-scale and multi-clustered settings. A cross-scale attention map is generated for human visual examination.

2.1 Feature Embedding and Phenotype Clustering

In the MIL community, most histopathological image analysis methods are divided into two stages [10,25]: (1) the self-supervised feature embedding stage and (2) the weakly

supervised feature-based learning stage. We follow a similar design that leverages our dataset to train a contrastive-learning model SimSiam [8] to extract high-level phenotype features from patches. All of the patches are then embedded into low-dimensional feature vectors for the classification in the second stage.

Inspired by [31], we implement K-means clustering to cluster patches on the patient level based on their self-supervised embeddings from the first stage since the high-level features are more comprehensive than low-resolution thumbnail images in representing phenotypes [33]. When gathering the patches equally from different clusters, the bag with the better generalization for the MIL model can be organized with distinctive phenotype patterns sparsely distributed on WSIs. In contrast, patches with similar high-level features can be aggregated for classification without spatial limitation.

2.2 Cross-Scale Attention Mechanism

We implement the MI-FCN encoder from DeepAttnMISL [31] as the backbone to encode patch embeddings from corresponding phenotype clusters and aggregate the instance-wise features to the patient-wise classification, which showed superior performance on survival prediction on WSIs. In the MIL community, several attention mechanisms [17,22] have been proposed for instance-relationship between different locations on WSIs. However, those methods are not aware of modeling multi-scale patterns from the pyramid-structured WSIs. Some approaches [15,20] have aggregated multi-scale features into deep learning models from WSIs. Unfortunately, those methods fail to exploit relationships between multiple resolutions at the same location.

To address this issue, we propose a cross-scale attention mechanism to represent distinctive awareness at different scales in the backbone. After separately encoding embedding features at different scales, the cross-scale attention mechanism from those encoding features is leveraged to consider the importance of each scale when aggregating multi-scale features at the same location. These attention scores are multiplied by representations from multiple scales to fuse the cross-scale embedding. The multi-scale representation F can be calculated by:

$$F = \sum_{s=1}^{S} a_s f_s \tag{1}$$

where

$$a_s = \frac{\exp \mathbf{W}^{\mathrm{T}} tanh(\mathbf{V} f_s^{\mathrm{T}})}{\sum_{s=1}^{S} \exp \mathbf{W}^{\mathrm{T}} tanh(\mathbf{V} f_s^{\mathrm{T}})} \tag{2}$$

$\mathbf{W} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times M}$ are trainable parameters in the cross-scale attention layer. L is the size of the MI-FCN output f_s, M is the output channel of the hidden layer in the cross-scale attention layer. Tangent element-wise non-linearity activation function $tanh(.)$ is implemented both negative and positive values for proper gradient flow. S is the number of the scales on WSIs. The attention-based instance-level pooling operator from [31] is then deployed to achieve patient-wise classification with cross-scale embedding.

2.3 Cross-Scale Attention Visualisation

The cross-scale attention maps from the cross-scale attention mechanism on WSIs are presented to show the distinctive contribution of phenotype features at different scales. The cross-scale attentions are mapped from patch scores of the cross-scale attention mechanism on WSIs, demonstrating the importance at multiple resolutions. This attention maps concatenate scale knowledge and location information can expand clinical clues for disease-guiding and exploration in different contexts.

3 Experiments

3.1 Data

50 H&E-stained Ascending Colon (AC) biopsies from [4], which are representative in CD, were collected from 20 CD patients and 30 healthy controls for training. The stained tissues were scanned at 20× magnification. For the pathological diagnosis, the 20 slides from CD patients were scored as normal, quiescent, mild, moderate, or severe. The remaining tissue slides from healthy controls were scored as normal. 116 AC biopsies were stained and scanned for testing with the same procedure as the above training set. The biopsies were acquired from 72 CD patients who have no overlap with the patients in the training data.

3.2 Experimental Setting

256×256 pixels patches were tiled at three scales (20×, 10× and 5×). For 20× patches, each pixel is equal to 0.5 Micron. Three individual models following the official SimSiam with a ResNet-50 backbone were trained at three scales, respectively. All three models were trained in 200 epochs with a batch size of 128 with the official setting. 2048-channel embedding vectors were received for all patches. K-means clustering with a class number of 8 was implemented to receive phenotype clustering within the single-scale features at three resolutions, and multi-scale features that include all resolutions for each patient.

10 data splits were randomly organized following the leave-one-out strategy in the training dataset, while the testing dataset was separated into 10 splits with a balanced class distribution. Each bag for MIL models was collected for each patient, equally selecting from different phenotype clustering classes, marked with a slide-wise label from clinicians. Negative Log-Likelihood Loss (NLLLoss) [30] was used to compare the slide-wise prediction for the bag with the weakly label. The validation loss was used to select the optimal model on each data split, while the mean value of the performance on 10 data splits was evaluated as the testing results. Receiver Operating Characteristic (ROC) curves with Area under the Curve (AUC) scores, Precision-Recall (PR) curves with Average Precision (AP) scores, and classification accuracy were used to estimate the performance of each model. We followed the previous work [12] to implement the bootstrapped two-tailed test and the DeLong test to compare the performance between the different models. The cross-scale attention scores were normalized within every single scale between 0 to 1.

Table 1. Classification performance on testing dataset.

Model	Patch scale	Clustering scale	AUC	AP	Acc
DeepAttnMISL(20×) [31]	Single	20×	0.7961	0.6764	0.7156
DeepAttnMISL(10×) [31]	Single	10×	0.7992	0.7426	0.6897
DeepAttnMISL(5×) [31]	Single	5×	0.8390	0.7481	0.7156
Gated attention [17]	Multiple	Multiple	0.8479	0.7857	0.7500
DeepAttnMISL [31]	Multiple	Multiple	0.8340	0.7701	0.7069
MDMIL-CNN [15]	Multiple	5×	0.8813	0.8584	0.7759
DSMIL [20]	Multiple	5×	0.8759	0.8440	0.7672
CS-MIL(Ours)	Multiple	5×	**0.8924**	**0.8724**	**0.8017**

Table 2. The bootstrapped two-tailed test and the DeLong test between different methods.

Model	p-value of AUC	p-value of AP
DeepAttnMISL(20×) [31]	0.004	0.001
DeepAttnMISL(10×) [31]	0.001	0.002
DeepAttnMISL(5×) [31]	0.048	0.004
Gated attention [17]	0.070	0.031
DeepAttnMISL [31]	0.009	0.002
MDMIL-CNN [15]	0.466	0.457
DSMIL [20]	0.350	0.201
CS-MIL(Ours)	Ref.	Ref.

4 Results

4.1 Performance on Classification

We implemented multiple DeepAttnMISL [31] models with patches at different scales with a single-scale setting. At the same time, we trained the Gated Attention (GA) model [17] and DeepAttnMISL model with multi-scale patches, without differentiating scale information. Patches from multiple scales are treated as instances when processing phenotype clustering and patch selection for MIL bags. Furthermore, we adopted a multi-scale feature aggregations, jointly adding embedding features from the same location at different scales into each MIL bag as [15]. We also concatenated embedding features from the same location at different scales as [20]. We followed above multi-scale aggregation to input phenotype features into the DeepAttnMISL backbone to evaluate the baseline multi-scale MIL models as well as our proposed method. All of the models were trained and validated within the same hyper-parameter setting and data splits.

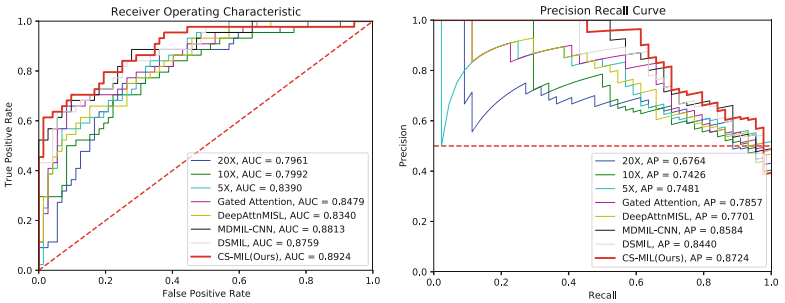

Fig. 3. ROC curves with AUC scores and PR curves with AP scores. This figure shows the ROC curves and PR curves of baseline models as well as the AUC scores and AP scores. The proposed model with cross-scale attention mechanism achieved superior performance in two metrics.

Table 3. Comparison of different cross-scale attention mechanism designs on testing dataset.

Id	Attention layer kernel	Activation function	AUC	AP	Mean of scores
1	Non-sharing	ReLU	0.8575	0.8559	0.8576
2	Non-sharing	Tanh	0.8848	0.8679	0.8763
3*	Sharing	ReLU	**0.8924**	**0.8724**	**0.8824**
4	Sharing	Tanh	0.8838	0.8609	0.8723

Testing Result. Table 1 and Fig. 3 indicates the performance of the performance while directly applying the models on the testing dataset in the CD classification task, without retraining. In general, single-scale models achieved worse performance compared to multi-scale models, indicating the benefit of external knowledge from multiple scale information. The proposed CS-MIL achieved better scores in all evaluation metrics, showing the benefits of the cross-scale attention which explores the inter-scale relationship at different scales in MIL. Table 2 shows the bootstrapped two-tailed test and the DeLong test to compare the performance between the different models.

Cross-Scale Attention Visualisation. Figure 4 represents cross-scale attention maps from the cross-scale attention mechanism on a CD WSI and normal WSI. The proposed CS-MIL can present distinctive importance-of-regions on WSIs at different scales, merging multi-scale and multi-region visualization. As a result, the 20× attention map highlights the chronic inflammatory infiltrates, while the 10× attention map focuses on

Table 4. Comparison of different bag sizes on testing dataset.

Bag size	AUC	AP	Mean of scores
64	0.8507	0.8220	0.8363
16	0.8690	0.8523	0.8606
08*	**0.8924**	**0.8724**	**0.8824**
01	0.8769	0.8261	0.8515

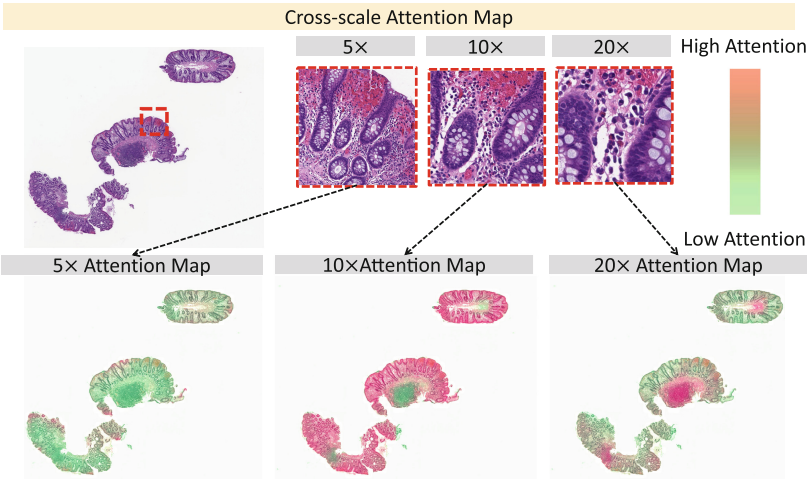

Fig. 4. Attention Map Visualization. This figure shows the cross-scale attention maps from the proposed model. The proposed CS-MIL can present importance-of-regions at different scales.

the crypt structures. Those regions of interest interpret the discriminative regions for CD diagnosis across multiple scales.

4.2 Ablation Studies

Inspired by [31] and [17], we estimated several attention mechanism designs in MIL with different activation functions. We formed the cross-scale attention learning into two strategies, differentiated by whether they shared the kernel weights while learning the embedding features from multiple scales. We also evaluated the performance of different bag sizes. As a result, as shown in Table 3, sharing the kernel weight for cross-scale attention learning with ReLU [3] achieved better performances with a higher mean value of multiple metrics. Table 4 demonstrates that a bag size of 8 is an optimal hyper-parameter for this study. The * is the proposed design.

5 Conclusion

In this work, we propose the addition of a cross-scale attention mechanism to an attention-guided MIL to combine multi-scale features with inter-scale knowledge. The inter-scale relationship provides extra knowledge of tissues-of-interest in lesions for clinical examination on WSIs to improve the CD diagnosis performance. The cross-scale attention visualization represents automatic scale-awareness and distinctive contributions to disease diagnosis in MIL when learning the phenotype features at different scales in different regions, offering an external AI-based clue for multi-scale pathological image analysis.

Acknowledgements. This work is supported by Leona M. and Harry B. Helmsley Charitable Trust grant G-1903-03793, NSF CAREER 1452485, and Veterans Affairs Merit Review grants I01BX004366 and I01CX002171, and R01DK103831.

References

1. AbdulJabbar, K., et al.: Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. Nat. Med. **26**(7), 1054–1062 (2020)
2. Abousamra, S., et al.: Multi-class cell detection using spatial context representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4005–4014 (2021)
3. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018)
4. Bao, S., et al.: A cross-platform informatics system for the gut cell atlas: integrating clinical, anatomical and histological data. In: Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications, vol. 11601, pp. 8–15. SPIE (2021)
5. Bejnordi, B.E., Litjens, G., Hermsen, M., Karssemeijer, N., van der Laak, J.A.: A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In: Medical Imaging 2015: Digital Pathology, vol. 9420, pp. 99–104. SPIE (2015)
6. Bejnordi, B.E., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22), 2199–2210 (2017)
7. Chen, J., Cheung, H.M.C., Milot, L., Martel, A.L.: AMINN: autoencoder-based multiple instance neural network improves outcome prediction in multifocal liver metastases. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 752–761. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_72
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021)
9. Con, D., van Langenberg, D.R., Vasudevan, A.: Deep learning vs conventional learning algorithms for clinical prediction in Crohn's disease: a proof-of-concept study. World J. Gastroenterol. **27**(38), 6476 (2021)
10. Dehaene, O., Camara, A., Moindrot, O., de Lavergne, A., Courtiol, P.: Self-supervision closes the gap between weak and strong supervision in histology. arXiv preprint arXiv:2012.03583 (2020)
11. Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep learning for whole slide image analysis: an overview. Front. Med. **6**, 264 (2019)
12. Gao, R., et al.: Cancer risk estimation combining lung screening CT with clinical data elements. Radiol. Artif. Intell. **3**(6), e210032 (2021)
13. Gao, Y., et al.: Multi-scale learning based segmentation of glands in digital colonrectal pathology images. In: Medical Imaging 2016: Digital Pathology, vol. 9791, pp. 175–180. SPIE (2016)
14. Gubatan, J., Levitte, S., Patel, A., Balabanis, T., Wei, M.T., Sinha, S.R.: Artificial intelligence applications in inflammatory bowel disease: emerging technologies and future directions. World J. Gastroenterol. **27**(17), 1920 (2021)
15. Hashimoto, N., et al.: 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 (CVPR) (2020)

16. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: 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 (2016)
17. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
18. Kiyokawa, H., et al.: Deep learning analysis of histologic images from intestinal specimen reveals adipocyte shrinkage and mast cell infiltration to predict postoperative Crohn disease. Am. J. Pathol. **192**, 904–916 (2022)
19. Kraszewski, S., Szczurek, W., Szymczak, J., Reguła, M., Neubauer, K.: Machine learning prediction model for inflammatory bowel disease based on laboratory markers. working model in a discovery cohort study. J. Clin. Med. **10**(20), 4745 (2021)
20. 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 (CVPR), pp. 14318–14328 (2021)
21. Lu, M.Y., et al.: Ai-based pathology predicts origins for cancers of unknown primary. Nature **594**(7861), 106–110 (2021)
22. 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. Nat. Biomed. Eng. **5**(6), 555–570 (2021)
23. Maksoud, S., Zhao, K., Hobson, P., Jennings, A., Lovell, B.C.: Sos: selective objective switch for rapid immunofluorescence whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3862–3871 (2020)
24. Mousavi, H.S., Monga, V., Rao, G., Rao, A.U.: Automated discrimination of lower and higher grade gliomas based on histopathological image analysis. J. Pathol. Inf. **6**(1), 15 (2015)
25. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H.M., Teuwen, J.: Deepsmile: self-supervised heterogeneity-aware multiple instance learning for dna damage response defect classification directly from h&e whole-slide images. arXiv preprint arXiv:2107.09405 (2021)
26. Skrede, O., et al.: Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet **395**(10221), 350–360 (2020)
27. Syed, S., Stidham, R.W.: Potential for standardization and automation for pathology and endoscopy in inflammatory bowel disease. Inflamm. Bowel Dis. **26**(10), 1490–1497 (2020)
28. Tokunaga, H., Teramoto, Y., Yoshizawa, A., Bise, R.: Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12597–12606 (2019)
29. Wang, S., et al.: RMDL: recalibrated multi-instance deep learning for whole slide gastric image classification. Med. Image Anal. **58**, 101549 (2019)
30. Yao, H., Zhu, D., Jiang, B., Yu, P.: Negative log likelihood ratio loss for deep neural network classification. In: Arai, K., Bhatia, R., Kapoor, S. (eds.) FTC 2019. AISC, vol. 1069, pp. 276–282. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-32520-6_22
31. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J.: Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med. Image Anal. **65**, 101789 (2020)
32. Yeshi, K., Ruscher, R., Hunter, L., Daly, N.L., Loukas, A., Wangchuk, P.: Revisiting inflammatory bowel disease: pathology, treatments, challenges and emerging therapeutics including drug leads from natural products. J. Clin. Med. **9**(5), 1273 (2020)
33. Zhu, X., Yao, J., Zhu, F., Huang, J.: Wsisa: Making survival prediction from whole slide histopathological images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7234–7242 (2017)