DSCENet: Dynamic Screening and Clinical-Enhanced Multimodal Fusion for MPNs Subtype Classification

Yuan Zhang[1 [0000−0002−0762−7514]], Yaolei Qi[1 [0000−0002−8531−7386]], Xiaoming Qi[1 [0000−0002−3238−2002]], Yongyue Wei[2 [0000−0002−1132−1796]], and Guanyu Yang[1,3 [0000−0003−3704−1722]] (✉)

[1] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
yang.list@seu.edu.cn
[2] Center for Public Health and Epidemic Preparedness Response, Peking University
[3] Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, Southeast University

Abstract. The precise subtype classification of myeloproliferative neoplasms (MPNs) based on multimodal information, which assists clinicians in diagnosis and long-term treatment plans, is of great clinical significance. However, it remains a great challenging task due to the lack of diagnostic representativeness for local patches and the absence of diagnostic-relevant features from a single modality. In this paper, we propose a Dynamic Screening and Clinical-Enhanced Network (DSCENet) for the subtype classification of MPNs on the multimodal fusion of whole slide images (WSIs) and clinical information. (1) A dynamic screening module is proposed to flexibly adapt the feature learning of local patches, reducing the interference of irrelevant features and enhancing their diagnostic representativeness. (2) A clinical-enhanced fusion module is proposed to integrate clinical indicators to explore complementary features across modalities, providing comprehensive diagnostic information. Our approach has been validated on the real clinical data, achieving an increase of 7.91% AUC and 16.89% accuracy compared with the previous state-of-the-art (SOTA) methods. The code is available at https://github.com/yuanzhang7/DSCENet.

Keywords: Computational pathology · Multimodal fusion · MPNs subtype classification

1 Introduction

The precise subtype classification of MPNs based on multimodal information is crucial [2]. MPNs are a heterogeneous group of hematologic malignancies with marked disparities in the progression of distinct subtypes, including polycythemia vera (PV), essential thrombocythemia (ET), prefibrotic/early primary myelofibrosis (PrePMF), and primary myelofibrosis (PMF) [1,5]. For example,

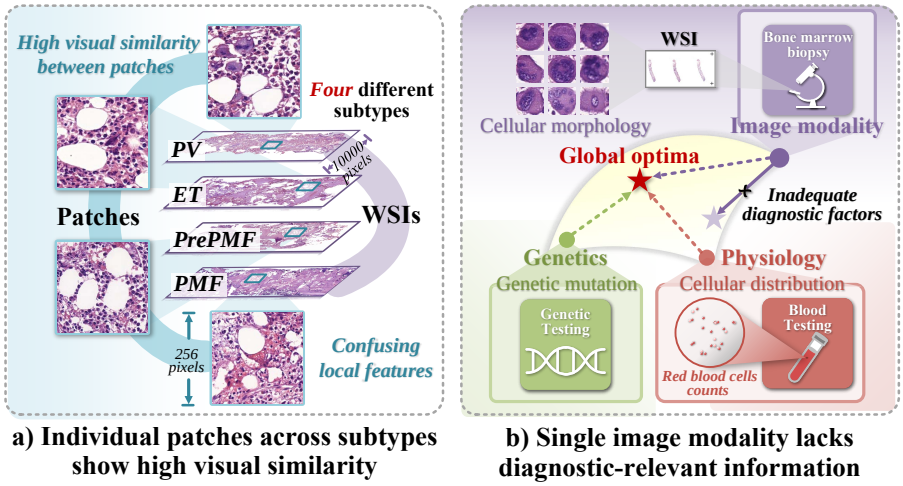

Fig. 1. a) Challenge 1: Individual patches show high visual similarity across four subtypes, leading to confusing local features. b) Challenge 2: Single image modality only provides morphological information, lacking other diagnostic-related features.

PMF has an increased risk for acute myelogenous leukemia transformation [27], while ET and PV give rise to thrombotic and bleeding complications [25]. However, relying solely on one modality, such as pathological images, yields limited insights [4]. The misclassification of subtypes may lead to inappropriate therapy, affecting treatment efficacy and patient survival [1]. Hence, accurate multimodal MPNs classification holds urgent clinical demands.

Pathology analysis on bone marrow biopsy whole slide images (WSIs) serves as the clinical gold standard for MPNs diagnosis [21,2], posing distinct challenges due to a unique paradigm of proliferative disorders. **Challenge 1: Individual patches lack diagnostic representativeness.** Due to the dispersed distribution and dense proliferation of blood cells, MPNs lack fixed lesion areas, causing individual patches to exhibit high visual similarity across four subtypes as Fig. 1. a) shown. Constraining tens of thousands of pixels of the WSI with a single category label is inadequate for capturing the intricate features and diversity inherent. The difficulty arises when local patches with similar textures struggle to represent different categories in the feature space, potentially leading to increased ambiguity and failure to encapsulate representative features. Hence, the effective selection of local patches from the perspective of holistic semantics is crucial for enhancing the model's ability to identify diagnostically relevant features. **Challenge 2: Single image modality lacks diagnostic-relevant features.** The diagnosis of MPNs relies on the comprehensive consideration of multi-modal diagnostic information such as morphology, physiology, and genetics [2]. However, the single modality of images only focuses on morphological features and lacks other diagnostic-relevant features, such as the driver mutations

from genetics and cell distributions from physiology, leading to the model being trapped in local optima as Fig. 1. b) shown. Therefore, incorporating clinical information guides the model in better learning the correlations and complementarities among different modalities, improving classification performance. Overall, the aforementioned MPN-related studies have feature extraction in isolation and have not facilitated multimodal fusion from the feature learning process.

Recently, while deep learning algorithms have demonstrated outstanding capabilities in pathological image diagnosis [12,29], the classification of hematological disorders like MPNs is still in its early stages [10]. Existing methods can be broadly classified into three types: 1) Pathology image-based approach, which often necessitates experts to perform cell-level annotations for the morphological analysis[16,14]. [6,23] extract features from WSIs based on the fingerprinting of megakaryocytes, requiring extensive cell annotation. [28] employs CNN models to classify patches directly but overlooks global contextual information. 2) Clinical information-based approach, which relies on indicators such as cell counts for diagnosis. [20] focuses on clinical indicators but faces limitations from the quality and completeness of available data. 3) Multimodal integration approach, which combines clinical data with pathology images. [26] integrates predictive probabilities from images and clinical data using statistical regression, but lacks interactive fusion at the feature learning level.

Multimodal fusion enhances comprehensive diagnostic information for MPNs, guiding the model to simultaneously learn features from different modalities, thereby exploiting cross-modal complementarity and improving classification accuracy [24,9]. To address challenge 1, we propose a dynamic screening (DS) module to filter the feature relevance of local patches, constraining the potential distribution differences to calibrate features, thereby boosting the model's representation. To address challenge 2, we propose a clinical-enhanced fusion (CF) module which introduces clinical information as additional guidance for image features learning. Given the dimensional disparity between modalities, our CF explores complementary features relevant to diagnosis in cross-modal learning, facilitating comprehensive diagnosis.

In this paper, we propose the multimodal fusion framework, dynamic screening and clinical-enhanced network (DSCENet), aiming at MPNs subtype classification. DSCENet capably explores complementary information across modalities, heightening the model's representative capacity and diagnostic accuracy. Our contributions are as follows:

- To the best of our knowledge, DSCENet achieves MPNs subtype classification based on multimodal fusion for the first time, obtaining significant improvements and contributing to reliable treatment plans.
- We propose a dynamic screening module to flexibly select the local patches, aiding the model in capturing pivotal features and improving representation capacity and generalization performance.
- We propose a novel clinical-enhanced fusion module to mine the diagnostic-relevant complementary features across modalities by using clinical features as additional guidance, facilitating comprehensive diagnostic accuracy.

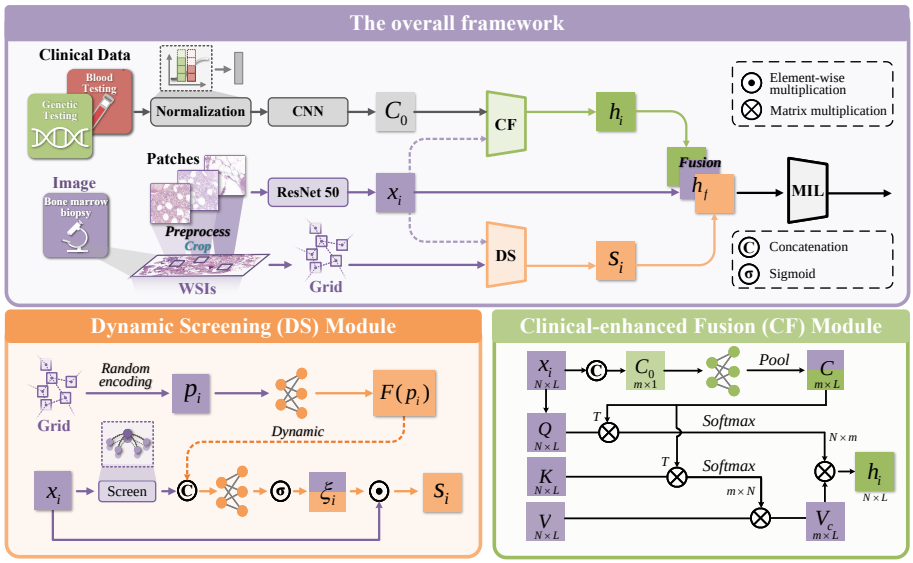

Fig. 2. The Framework of our DSCENet. The DS module aims to dynamically screen the local patches for better feature representation and clinical-enhanced fusion module explores the complementary diagnostic information.

– Our DSCENet outperforms the SOTA methods on the real clinical dataset consisting of 383 cases with an accuracy of 83.12% and an AUC of 96.43%.

2 Methodology

Fig. 2 illustrates the overall framework of our method to integrate WSIs and clinical data, which consists of two main modules. The dynamic screening module aims to flexibly screen the feature learning of local patches, improving the representative capacity of WSIs. The clinical-enhanced fusion module intends to integrate clinical indicators to guide the exploration of complementary features across modalities, offering comprehensive diagnostic information.

2.1 Dynamic Screening Module

Dynamic random encoding of the grid. Our approach is dedicated to enhancing the representation of local patches in the holistic context of WSIs, i.e., $B = \{x_i\}$ denotes the bag of features per WSI, where i is the index of the cropped patches. The feature embedding of patch $x_i \in \mathbb{R}^{N\times L}$ is extracted by the pre-trained ResNet 50 [18], where N is the number of patches and L is the feature dimension. Instead of the absolute position encoding, we dynamically random encode the sequences of patches as the random grid $p_i \in \mathbb{R}^N$. Then, two

fully-connected layers with ReLU activation $\mathcal{F}(\cdot)$ are used to realize dynamic enhancement encoding. The dynamic random encoding of the grid $\mathcal{F}(p_i)$ efficiently adapts to the varying number of patches within each WSI, thereby aiding the model in capturing the overall structure and patterns in the sequence.

Holistic screening of patch feature. Unlike the top-ranking tiles [7] that discards subsequent patches, the dispersed distribution of MPNs requires retraining the holistic feature information through dynamic screening. To capture the holistic semantics, we also perform global averaging $\mathcal{G}(\cdot)$ on $x_{i,j}$ to save the computational cost. Following by a full-connection layer and the gating sigmoid activation for augmenting embedding by the textural knowledge, the selection descriptor of patch feature ξ_i is then given by:

$$\xi_i = sigmoid(\mathcal{G}(x_i) \oplus \mathcal{F}(p_i)), \tag{1}$$

where $\oplus$ denotes concatenation.

Finally, we also implement a residual design to alleviate the adverse impacts of inaccurate selections. The dynamic selected feature s_i is then given by:

$$s_i = x_i \otimes \xi_i + x_i, \tag{2}$$

where $\otimes$ denotes pixel-wise multiplication. As the selection descriptor s_i are mutually independent and differ among local patches, they generate different augmented versions to calibrate the feature representation learning and eventually adapt to the holistic semantics.

2.2 Clinical-enhanced Fusion Module

Given the dimensional disparity between WSIs and clinical data in MPNs diagnosis, multimodal fusion methods like alignment are suboptimal [9]. Inspired by [11], which applies agent tokens to enhance efficiency, we design the clinical-enhanced fusion module to mine the diagnostic-relevant complementary features by introducing clinical features as additional guidance between clinical-enhanced query and clinical-enhanced value block.

Clinical query block. Clinical indicators $C_o \in \mathbb{R}^M$ includes genetics, reflecting the correlation between gene mutations and specific subtypes [3], and physiology, describing dynamic physiological metabolic distributions [2], where m is the number of clinical indicators. Clinical-enhanced query $C \in \mathbb{R}^{M \times L}$ is derived from the concatenation of clinical feature and image feature followed by fully connected layers and ReLU activation and scale pooling. The image feature $Q = x_i W_Q, K = x_i W_K, V = x_i W_V$, where $W_{Q/K/V}$ denote projection metrices and d_k is the dimension. Clinical query block performs attention calculations between C, K, V to query the focus of clinical attention from images (Equ. 3).

$$V_c = Attention(C, K, V) = softmax(\frac{CK^T}{\sqrt{d_k}})V, \tag{3}$$

$$h_i = Attention(Q, C, V_c) = softmax(\frac{QC^T}{\sqrt{d_k}})V_c, \tag{4}$$

Clinical-enhanced shared block. To obtain complementary features h_i cross modalities, it is essential to perform the second attention with image feature as Equ. 4. Clinical-enhanced shared block achieves bidirectional attention in the shared feature space of both image and clinical modalities, offering comprehensive diagnostic insights and ultimately enhancing diagnostic precision.

Finally, the comprehensive feature h_f is obtained after combining dynamical screening and clinical-enhanced fused feature: $h_f = s_i \oplus h_i$. Then we adopt the widely-used multiple instance classifier [19] and Softmax for classification.

3 Experiments

3.1 Experimental Settings

Dataset. To validate the effectiveness of our method, we conducted a multicenter study on MPNs data. 383 cases of MPNs are collected from 3 independent medical centers, including 81 PV, 126 ET, 88 PrePMF, and 88 PMF cases. Each patient tasks one WSI paired with its corresponding clinical information with no missing data, including gender, age, and blood test report (hemoglobin, white blood cell, red cell mass, hematocrit, platelet count) and genetic mutations (JAK2, MPL, CALR). The WSIs are stained by Hematoxylin and Eosin and scanned at 0.2 m/pixel using PANNORAMIC SCAN 150.

Implementation Details. We preprocess WSIs into non-overlapping patches measuring 256×256 pixels at $20\times$ magnification and use a pre-trained ResNet-50 [18] to extract the 1024-dimensional morphological features. All experiments are conducted on the two NVIDIA GeForce RTX 4090 (24 GB). The datasets are randomly split into training, validation, and testing sets with a ratio of 6: 2: 2. A total of 200 training epochs are conducted with a learning rate of 1×10^{-4} using the Adam optimizer and cross-entropy loss.

Evaluation Metric. We perform comparative experiments and ablation studies to demonstrate the advantages of our methods. The area under the curve (AUC), Receiver Operating Characteristic curves (ROC), accuracy(ACC), precision, recall, and F1 score (F1) are employed to evaluate the performance.

3.2 Result and Analysis

We performed comparative experiments and ablation studies to demonstrate the advantages of our proposed framework. The classical classification network CLAM [18], Transmil [22], and DSMIL [17] are compared to validate the accuracy for the single image modality. We also compare with SNN [15] for the single modality of clinical data. To validate the performance of multimodal fusion, we compared the AMIL [13], MCAT [8] and CMTA [30]. These models are trained on the same dataset and evaluated by the above metrics. For a fair comparison, the same CNN extractor and cross-entropy loss are adopted for all methods.

Table 1. Comparisons between our proposed method and other SOTA approaches (%). *concat* is concatenation fusion and *bilinear* is bilinear pooling fusion.

	WSIs	Clinical data	ACC	AUC	Precision	Recall	F1
CLAM_SB [18]	✓		32.47	51.30	8.12	25.00	12.25
CLAM_MB [18]	✓		64.94	83.36	64.13	63.10	63.23
DSMIL [17]	✓		53.25	78.86	72.66	60.84	54.95
Transmil [22]	✓		64.94	85.52	62.73	62.36	62.00
SNN [15]		✓	63.64	82.61	48.34	63.44	54.80
AMIL_bilinear [13]	✓	✓	48.05	76.03	50.79	46.99	45.40
AMIL_concat [13]	✓	✓	58.44	88.52	53.02	64.93	56.21
CMTA [30]	✓	✓	58.44	87.19	62.96	62.38	57.13
MCAT [8]	✓	✓	66.23	86.28	63.80	66.83	64.58
Ours	✓	✓	**83.12**	**96.43**	**83.32**	**83.15**	**83.02**

Quantitative Evaluation. In the quantitative evaluation, our method achieve the best classification performance with 83.12% accuracy, 96.73% AUC, 83.32% precision, 83.15% recall, and 83.02% F1 score in the Table 1. Our method outperforms the second-best fusion model (MCAT) by 16.89% in AUC and 7.91% in accuracy and achieves an 18.18% higher accuracy and 10.91% higher AUC compared to the single-image modality model (Transmil). On the one hand, single-modality approaches utilizing either clinical information or WSIs have demonstrated moderate classification performance, indicating that individual modalities indeed contain specific diagnostic-relevant features. On the other hand, other modality fusion methods ignore the huge dimensional difference or design for prognostic tasks, which may not be applicable in the unique diagnostic scenarios of MPNs, thus failing to achieve the ideal alignment between two image and non-image feature spaces. However, our method employs a monitoring process augmented with additional clinical information, aiding the network in learning better features and improving classification performance.

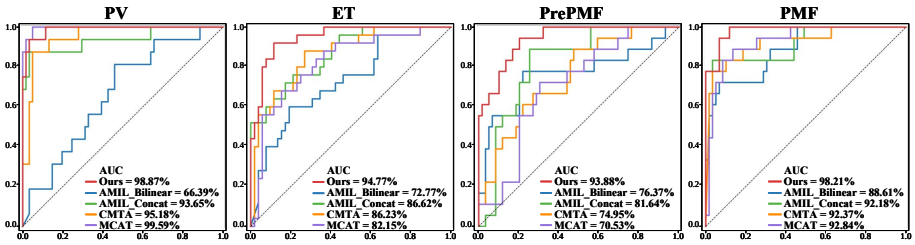

Fig. 3. ROC curves for different methods on PV, ET, PrePMF, and PMF.

Table 2. Ablation results of our method (%). DS is the dynamic screening module, and CF is the clinical-enhanced fusion module.

	ACC	AUC	Precision	Recall	F1
w/o DS, w/o CF	63.64	81.56	72.66	60.84	54.95
w/o CF	76.62	92.51	77.28	77.16	74.95
w/o DS	76.62	93.90	75.35	76.38	75.09
Ours	**83.12**	**96.43**	**83.32**	**83.15**	**83.02**

Qualitative Evaluation. In the subtype classification task of MPNs, our model performs significantly superior for four subtypes on the ROC curves. The AUC values of our model are 98.87%, 94.77%, 93.88%, and 98.21% for PV, ET, PrePMF, and PMF, respectively, as depicted in Fig. 3. Our model also surpasses the second-highest model by 8.54%, 18.93%, and 5.84% in terms of AUC for ET, PrePMF, and PMF, respectively. Our model demonstrates well-balanced classification performance and achieves the top AUC value of 96.43%, indicating effective fusion of clinical data with images at the cross-modal feature space, contributing to complementary diagnostic information.

Ablation Study. We evaluated the importance of monitoring and fusion modules by ablation experiments (Table 2). The DS module brings significant improvement for classification, suggesting effective filtering of redundant features from patches. The CF module also improved the AUC by 12.34%, indicating that clinical data with images enhances the representation of diagnostic relevance.

Model analysis of confusion matrices. The confusion matrices of our model on the testing dataset of 77 samples demonstrate the distinguished classification performance of our model, achieving an accuracy of 83.11% in Fig. 4. It is evident that other methods are prone to misclassify ET as PrePMF or PV, whereas our method improves the accuracy of ET while maintaining the correctness of others. ET has dense cellular features, suggesting that our method can effectively enhance the local features with stronger diagnostic discrimination.

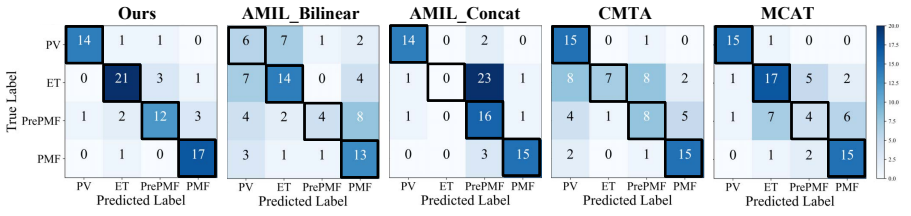

Fig. 4. Confusion matrices on the testing set, with true labels on the vertical axis and predictions on the horizontal axis. The deeper colors indicate higher accuracy.

4 Conclusion

We propose a DSCEnet for the multimodal subtype classification of MPNs for the first time. This method effectively integrates pathological images and clinical data, achieving the optimal classification performance. Furthermore, we utilize a dynamic screening module to select patch features for better representation and a clinical-enhanced fusion module to explore diagnostic-relevant features for comprehensive diagnoses. We have pioneered a promising future for multimodal diagnosis of MPNs, offering vast prospects in diagnosis and treatment.

Acknowledgments. We thank the Big Data Computing Center of Southeast University for providing the facility support. We also thank Jiangsu Provincial People's Hospital, Peking Union Medical College Hospital, and the First Affiliated Hospital of Soochow University for providing the data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arber, D.A., Orazi, A., Hasserjian, R., et al.: The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. Blood, The Journal of the American Society of Hematology **127**(20), 2391–2405 (2016)
2. Arber, D.A., Orazi, A., Hasserjian, R.P., et al.: International consensus classification of myeloid neoplasms and acute leukemias: integrating morphologic, clinical, and genomic data. Blood, The Journal of the American Society of Hematology **140**(11), 1200–1228 (2022)
3. Barbui, T., Thiele, J., Gisslinger, H., et al.: The 2016 who classification and diagnostic criteria for myeloproliferative neoplasms: document summary and in-depth discussion. Blood cancer journal **8**(2), 15 (2018)
4. Barbui, T., Thiele, J., et al.: Myeloproliferative neoplasms: Morphology and clinical practice. American journal of hematology **91**(4), 430–433 (2016)
5. Baumeister, J., Chatain, N., et al.: Progression of myeloproliferative neoplasms (mpn): diagnostic and therapeutic perspectives. Cells **10**(12), 3551 (2021)
6. Brück, O.E., Lallukka-Brück, S.E., et al.: Machine learning of bone marrow histopathology identifies genetic and clinical determinants in patients with mds. Blood cancer discovery **2**(3), 238–249 (2021)
7. Campanella, G., Hanna, M.G., Geneslaw, L., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
8. Chen, R.J., Lu, M.Y., Weng, W.H., et al.: 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 (2021)
9. Cui, C., Yang, H., Wang, Y., et al.: Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering (2023)
10. Elsayed, B., et al.: Applications of artificial intelligence in philadelphia-negative myeloproliferative neoplasms. Diagnostics **13**(6), 1123 (2023)

11. Han, D., Ye, T., Han, Y., et al.: Agent attention: On the integration of softmax and linear attention. arXiv preprint arXiv:2312.08874 (2023)
12. He, Y., Huang, F., Jiang, X., et al.: Foundation model for advancing healthcare: Challenges, opportunities, and future directions (2024)
13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
14. Kimura, K., Ai, T., Horiuchi, Y., et al.: Automated diagnostic support system with deep learning algorithms for distinction of philadelphia chromosome-negative myeloproliferative neoplasms using peripheral blood specimen. Scientific Reports **11**(1), 3367 (2021)
15. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. Advances in neural information processing systems **30** (2017)
16. Krichevsky, S., Ouseph, M.M., Zhang, Y., et al.: A deep learning-based pathomics methodology for quantifying and characterizing nucleated cells in the bone marrow microenvironment. Blood **142**, 2294 (2023)
17. 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)
18. Lu, M.Y., Williamson, D.F., Chen, T.Y., et al.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
19. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems **10** (1997)
20. Meggendorfer, M., Walter, W., Haferlach, C., et al.: Deep learning algorithms support distinction of pv, pmf, and et based on clinical and genetic markers. Blood **130**, 4223 (2017)
21. Ryou, H., Lomas, O., Theissen, H., et al.: Quantitative interpretation of bone marrow biopsies in mpn—what's the point in a molecular age? British Journal of Haematology **203**(4), 523–535 (2023)
22. Shao, Z., Bian, H., Chen, Y., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
23. Sirinukunwattana, K., Aberdeen, A., Theissen, H., et al.: Artificial intelligence–based morphological fingerprinting of megakaryocytes: a new tool for assessing disease in mpn patients. Blood advances **4**(14), 3284–3294 (2020)
24. Song, A.H., Jaume, G., et al.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
25. Tefferi, A., Barbui, T.: Polycythemia vera and essential thrombocythemia: 2017 update on diagnosis, risk-stratification, and management. American journal of hematology **92**(1), 94–108 (2017)
26. Wang, R., Shi, Z., Zhang, Y., et al.: Development and validation of deep learning model for diagnosis and subtypes differentiation of myeloproliferative neoplasms using clinical data and digital pathology. Blood **142**, 123 (2023)
27. Yogarajah, M., Tefferi, A.: Leukemic transformation in myeloproliferative neoplasms: a literature review on risk, characteristics, and outcome. In: Mayo Clinic Proceedings. pp. 1118–1128. Elsevier (2017)
28. Yusof, U.K.M., Mashohor, S., et al.: Hyperparameter selection in deep learning model for classification of philadelphia-chromosome negative myeloproliferative

neoplasm. In: Proceedings of the 11th International Conference on Robotics, Vision, Signal Processing and Power Applications: Enhancing Research and Innovation through the Fourth Industrial Revolution. pp. 27–32. Springer (2022)
29. Zhang, Y., Qi, Y., et al.: Fedsoda: Federated cross-assessment and dynamic aggregation for histopathology segmentation. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1656–1660 (2024)
30. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21485–21494 (2023)