Enhancing Soft Tissue Sarcoma Classification by Mitigating Patient-Specific Bias in Whole Slide Images

Weiping Lin[1], Runchen Zhu[2], Wentai Hou[3], Jiacheng Wang[4] (✉), Yixuan Lin[1], Rui Chen[5], Na Ta[6], and Liansheng Wang[1](✉)

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China
{wplin, yixuanlin}@stu.xmu.edu.cn, lswang@xmu.edu.cn

[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
zhurunchen@stu.xmu.edu.cn

[3] The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming, China
houwentai@fudan.edu.cn

[4] Manteia Technology Co. Ltd, Xiamnen, China
jiachengw@stu.xmu.edu.cn

[5] Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
drchenrui@foxmail.com

[6] Shanghai Changhai Hospital, Shanghai, China
drtana1988@163.com

Abstract. Soft tissue sarcomas (STS) are a rare and heterogeneous group of malignant tumors that arise in soft tissues throughout the body. Accurate classification from whole slide images (WSIs) is essential for diagnosis and treatment planning. However, STS classification faces a significant challenge due to patient-specific biases, where WSIs from the same patient share confounding non-tumor-related features, such as anatomical site and demographic characteristics. These biases can lead models to learn spurious correlations, compromising their generalization. To address this issue, we propose a novel multiple instance learning framework that explicitly mitigates patient-specific biases from WSIs. Our method leverages supervised contrastive learning to extract patient-specific features and integrates a bias-mitigation strategy based on propensity score matching. Extensive experiments on two STS datasets demonstrate that our approach significantly improves classification performance. By mitigating patient-specific biases, our method improves the reliability and generalization of the model, contributing to a more accurate and clinically reliable STS classification. To facilitate direct clinical application and support decision-making, the code, trained models, and testing pipeline will be publicly available at https://github.com/Lanman-Z/MPSF.

Keywords: Soft tissue sarcomas · Patient-specific bias · Whole slide images.

1 Introduction

Soft tissue sarcomas (STS) are a rare and heterogeneous group of malignant tumors, accounting for approximately 1% of all adult malignancies [3]. STS can arise at various anatomical locations, including the extremities, trunk, retroperitoneum, and head and neck, without a predominant site. STS exhibits extensive histological and molecular heterogeneity, with over 50 recognized subtypes [15], such as rhabdomyosarcoma (RMS), undifferentiated pleomorphic sarcoma (UPS), leiomyosarcoma (LMS), liposarcoma (LPS), and synovial sarcoma (SS). Accurate subtype classification is essential for guiding treatment decisions and enabling personalized therapeutic strategies, as different STS subtypes show highly variable responses to chemotherapy, radiotherapy, and targeted therapies.

The diagnosis of STS remains highly challenging due to significant pathological overlap among subtypes and inter-observer variability, leading to potential misdiagnoses even among experts. Deep learning has shown promise in pathology image analysis, often surpassing human performance [16]. Several studies have explored its application in STS classification. For instance, Foersch et al. [4] used DenseNet121 to classify regions of interest (ROIs) from five major subtypes, while Tomohito et al. expanded the coverage to 11 subtypes [6]. Additionally, some studies have focused on specific subtypes, such as rhabdomyosarcoma [20,5,11], leiomyosarcoma [19], myxoid soft tissue sarcoma [18], and canine soft tissue sarcoma [12]. However, most existing methods either rely on manually annotated regions rather than WSI-level analysis or cover only a limited number of subtypes, restricting their clinical applicability.

Applying deep learning to STS subtype classification posed two major challenges. As shown in Fig. 1, STS can occur in various anatomical locations, meaning that WSIs may contain not only diagnostic features of STS but also information related to the tumor location. Second, different subtypes exhibit demographic differences, which may also be embedded in WSIs. Such additional information, termed patient-specific bias, is unrelated to the morphological feature of STS subtypes and offer little diagnostic value, yet they are inevitably present in WSIs. When these irrelevant features exhibit spurious correlations with subtype labels, models may rely on them for predictions rather than true diagnostic cues, leading to compromised performance and poor generalization. Thus, identifying and mitigating these biases is crucial for improving model generation and reliability. However, most existing studies overlook this issue, and common multiple instance learning (MIL) methods fail to explicitly address patient-level confounders, limiting their effectiveness in real-world STS classification.

To address this challenge, we propose a MIL framework that identifies and mitigates patient-specific biases from a casual perspective. Causal inference provides a powerful tool for analyzing spurious correlations and has been successfully applied in pathology image analysis [8,9,1]. First, we analyze the presence of irrelevant information in STS WSIs and identify patient-specific biases. Then, we design a supervised contrastive learning method to extract patient-specific features in WSIs. Finally, we inject these features into each training sample, making them indistinguishable across WSIs and preventing the model from relying

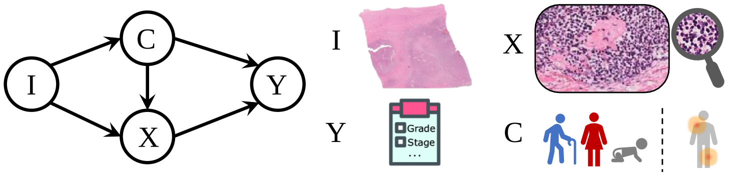

Fig. 1. Casual graph for STS classification. I and Y represent the WSI and prediction. X denotes the true pathological feature. C denotes the patient-specific feature.

on them for STS classification. Extensive experiments on our datasets demonstrate the effectiveness of our method. Our main contributions are summarized as follows: (1) We propose a novel model-agnostic MIL framework tailored to mitigate patient-specific biases, improving classification performance, generalization ability, and reliability. (2) We innovatively propose a supervised contrastive learning method for explicitly extract patient-specific information in WSIs. (3) We introduce a bias-mitigation strategy that integrates patient-specific features into each WSI, ensuring models focus on diagnostic-relevant patterns. (4) The code, trained models, and a testing pipeline for new samples will be released to support clinical decision making.

2 Method

We propose a novel MIL framework that mitigates patient-specific biases in WSIs, as shown in Fig. 2. First, we analyze images from a causal inference perspective to separate diagnostic-related features from confounding patient-specific information. Then, we introduce a supervised contrastive learning approach to extract and encode such information. To minimize their influence, we aggregate all patient-specific features into a confounder embedding, which is then integrated with the original patch feature embedding for classification.

2.1 STS Classification from Causal View

In this section, we analyze diagnostic-related and patient-specific features in STS WSIs. Diagnostic-related features include cell morphology, tissue architecture, and key pathological patterns essential for diagnosis. In contrast, patient-specific features refer to confounding information such as occurrence site and demographic characteristics. For example, if all WSIs from a LPS patient originate from the retroperitoneum, a model may incorrectly associate retroperitoneal features with LPS, leading to misclassification when encountering LPS in other anatomical sites. Similarly, demographic differences can also introduce biases. For instance, LMS predominantly affects middle-aged and elderly individuals, while RMS is more common in children [17]. These spurious correlations can mislead models, reducing generalization and reliability.

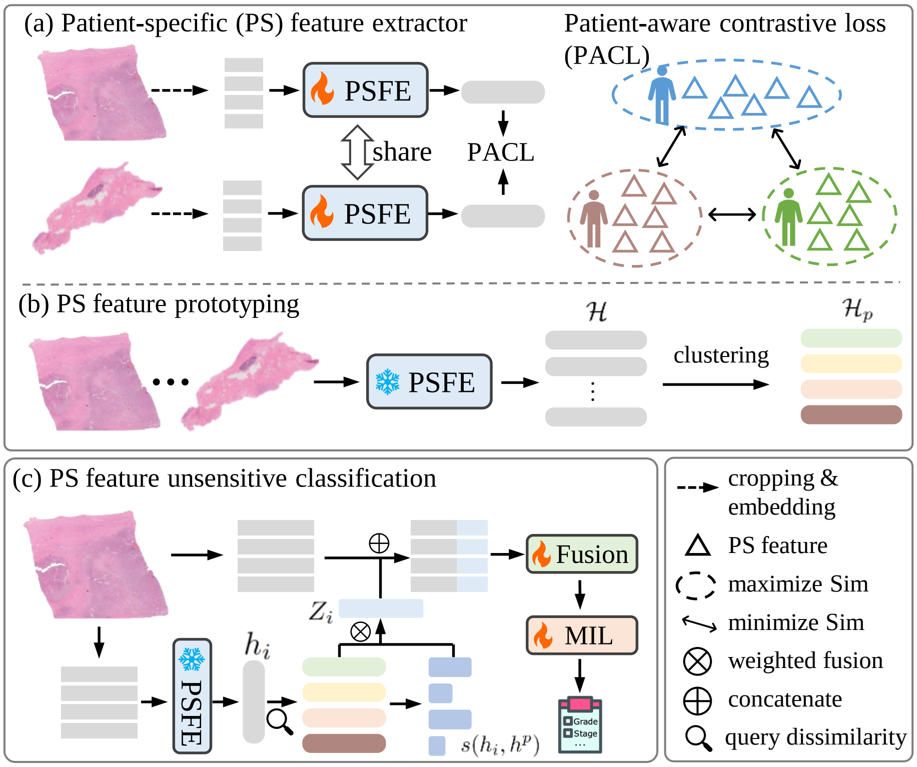

Fig. 2. Overview. PSFE is trained to extract patient-specific features from WSIs, which are summarized as prototypes. Each WSI is integrated with all PS prototypes to suppress the impact of PS feature.

We represent the above analysis using a causal graph [13] in Fig. 1, where C represents the confounder, namely patient-specific biases. X denotes the true causal factors, i.e., the diagnostic-related features of STS. Y represents the outcome, i.e., the STS diagnosis. $C \to Y$ represents a spurious correlation, which arises due to dataset-specific coincidences and lacks generalizability. In contrast, $X \to Y$ represents the true diagnostic process, where predictions are made based on pathological features, reflecting valid medical knowledge. Ideally, the model should minimize reliance on $C \to Y$ and instead focus on $X \to Y$.

2.2 Identifying Patient-Specific Information

The core idea of our method is to mitigate the impact of $C \to Y$. Therefore, the first step is to identify C from WSIs. C represents patient-level biases in STS WSIs, such as specific occurrence sites and demographic differences. These confounding features are often subtle and may not be visually discernible. However, regardless of their exact form, one thing is certain. WSIs from the same patient share similar patient-specific features, while those from different patients tend to exhibit distinct patient-specific features. Based on this principle, we leverage the patient-WSI association as a supervised signal and design a novel supervised contrastive learning method to extract patient-specific features.

Specifically, we construct positive and negative sample pairs based on the patient-WSI association. $D = \{(x_i, y_i, c_i) \mid 1 \le i \le N\}$ represents the dataset, where x_i is the WSI, y_i is the corresponding STS subtype label, and c_i denotes the patient to whom x_i belongs. WSIs from the same patient form positive pairs, while those from different patients form negative pairs. The sample pairs and labels are defined in Eq. 1, where $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the condition is satisfied and 0 otherwise. Since our goal is to extract only patient-specific information, we ensure that all selected sample pairs share the same STS label, preventing diagnostic features from interfering with this process.

$$\mathcal{B} = \{(x_i, x_j, y_{\mathcal{B}} = \mathbb{I}(c_i = c_j)) \mid y_i = y_j\} \tag{1}$$

We employ an MIL model as the feature extractor $\mathcal{M}(\cdot) \in \mathbb{R}^d$, with a constraint to ensure that the slide-level embeddings of positive sample pairs are as similar as possible. The loss function for training $\mathcal{M}(\cdot)$ is defined in Eq. 2, where $s(\cdot, \cdot)$ denotes the cosine similarity.

$$\mathcal{L} = -y_{\mathcal{B}} \log s(\mathcal{M}(x_j), \mathcal{M}(x_j)) - (1 - y_{\mathcal{B}}) \log(1 - s(\mathcal{M}(x_j), \mathcal{M}(x_j))) \tag{2}$$

2.3 Mitigating Patient-Specific Bias

Revisiting the causal graph, to reduce the interference of C, we enforce all samples to have similar C. In this way, during model training, C becomes insufficient for distinguishing different WSIs, forcing the model to focus on X. Specifically, we use $\mathcal{M}$ to extract patient-specific features from WSIs, i.e., $\mathcal{H} = \{h_i = \mathcal{M}(x_i) \mid x_i \in D_{train}\}$. We further cluster these patient-specific features to obtain k patient-specific feature prototypes $\mathcal{H}_p = \{h_i^p \mid 1 \le i \le k\}$. $\mathcal{H}_p$ represents the patient-specific features present in the dataset. During MIL model training for STS classification, x_i is input into $\mathcal{M}$ to obtain its patient-specific feature h_i. Since each x originally possesses different patient-specific features, we determine the weight of each prototype in $\mathcal{H}_p$ based on the dissimilarity between h_i and each prototype. The higher the dissimilarity, the higher the corresponding weight. We then perform a weighted average based on these weights and prototypes, resulting in a confounding embedding Z_i for x_i (Eq. 3). Finally, Z_i is concatenated with the patch embeddings of x_i, guiding the model to obtain more accurate attention scores and prediction.

$$Z_i = \sum_{h^p \in H_p} [1 - s(h_i, h^p)] \cdot h^p \tag{3}$$

Table 1. Performance comparison of MIL models with and without our framework. Models were trained on the training set of SARC-1 and tested on the testing set of SARC-1 and all samples from SARC-2.

		SARC-1			SARC-2		
		AUC	ACC	F1-score	AUC	ACC	F1-score
ABMIL	w/o ours	0.8962	0.7598	0.5668	0.8490	0.6730	0.6280
	w/ ours	0.9164	0.7832	0.6506	0.8800	0.7280	0.6490
	Δ	0.0202	0.0234	0.0838	0.0310	0.0550	0.0210
CLAM	w/o ours	0.9112	0.7512	0.6228	0.8650	0.7170	0.6380
	w/ ours	0.9148	0.7630	0.6266	0.9020	0.7280	0.6620
	Δ	0.0036	0.0118	0.0038	0.0370	0.0110	0.0240
TransMIL	w/o ours	0.9016	0.7764	0.6122	0.8710	0.6850	0.6010
	w/ ours	0.9136	0.7856	0.6606	0.8910	0.7200	0.6460
	Δ	0.0120	0.0092	0.0484	0.0200	0.0350	0.0450

Table 2. The ablation study on the embedding dimension (d) of PS features and the fusion strategy.

	SARC-1			SARC-2		
	AUC	ACC	F1-score	AUC	ACC	F1-score
d=64	**0.9170**	0.7712	0.5954	0.8650	0.6810	0.6300
d=128	0.9164	**0.7832**	**0.6506**	**0.8800**	0.7280	**0.6490**
d=256	0.9164	0.7718	0.6014	0.8530	0.6730	0.6020
d=512	0.9118	0.7592	0.5448	0.8570	0.6810	0.6120
w/o prototyping	0.9000	0.7776	0.6194	0.8670	**0.7360**	0.6416
Opposite Fusion	0.9022	0.7652	0.5822	0.8440	0.6770	0.5760
Direct Concatenation	0.9094	0.7720	0.6284	0.8730	0.6978	0.6312

3 Experiments

Datasets. We collected the SARC-1 STS dataset, comprising 604 WSIs from 179 patients, covering five common STS subtypes: LPS (214), LMS (260), RMS (50), SS (20), and UPS (60). Additionally, we obtained the SARC-2 dataset from another hospital for external validation, which contains 254 WSIs.

Implementation Details. The patient-specific feature extractor is also an MIL model, with the same architecture as the MIL model used for STS classification. We constructed 1,876 positive and negative sample pairs from SARC-1 for training the feature extractor, and 340 pairs for validation, with no patient overlap between the training and validation sets. The dimension of the patient-specific features is set to 128, and cosine similarity is used for all similarity measurements. The number of patient-specific feature prototypes is set to $k = 10$. The feature extractor used is UNI [2].

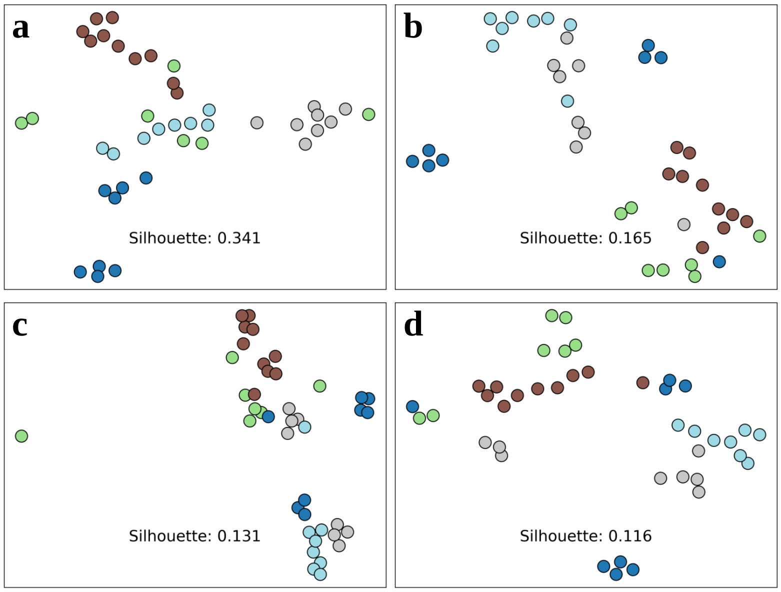

Fig. 3. TSNE visualization of slide-level embeddings. (a) the patient-specific feature extraction model, (b) the baseline model, (c) the baseline with our method but uses direct slide-level concatenation, and (d) the baseline with our method.

Performance Comparison. We evaluated our method by integrating it with ABMIL [7], CLAM [10], and TransMIL [14]. Performance was compared using AUC, accuracy (ACC), and F1-score. Models were trained with five-fold cross-validation on SARC-1 and tested on SARC-2 for external validation. As shown in Table 1, incorporating our framework consistently improved model performance across all metrics, demonstrating its effectiveness in reducing patient-specific biases and enhancing classification performance and generalization. The performance gains were more pronounced on SARC-2, highlighting its ability to mitigate spurious correlations and improve real-world applicability.

Ablation Study. First, we assess the impact of the patient-specific feature dimension by evaluating different sizes: 64, 128, 256, and 512. Second, we examine the effect of the fusion method defined in Eq. 3 by testing the following variations: (1) Removing the prototyping process in Fig. 2(b), where all patient-specific features are directly matched to the WSI without clustering into prototypes. (2) Reversing the fusion strategy, where higher dissimilarity between the patient-specific feature of WSIs and the prototypes in $\mathcal{H}$ results in a lower weight, opposite to our setting (Eq. 4). (3) Directly concatenating Z (Eq. 3) with the slide-level embedding of STS classification MIL.

$$Z_i = \sum_{h^p \in H_p} s(h_i, h^p) \cdot h^p \tag{4}$$

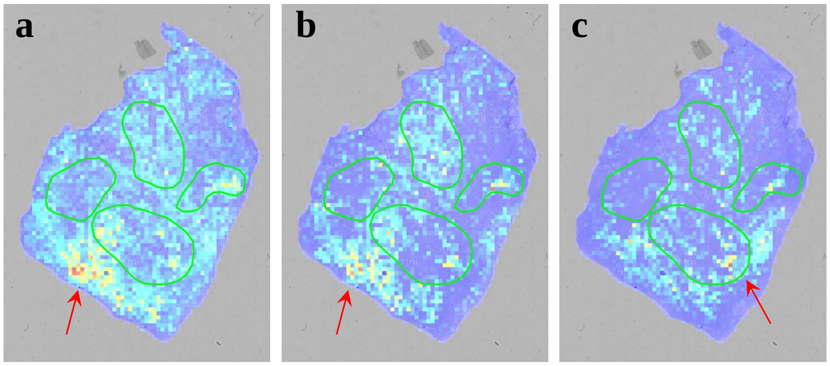

Fig. 4. Comparison of attention maps. (a) the baseline model, (b) the baseline with our method but uses direct slide-level concatenation, (c) the baseline with our method.

Table 2 presents the results, showing that the best performance is achieved when $d = 128$. The fusion strategy we utilize yields the best results.

Analysis of Patient-Specific Features. To evaluate the effectiveness of patient-specific features, we assessed the performance of the patient-specific feature extractor. During training, we used the SARC-1 dataset while holding out a subset of patients. Following Eq. 1, we constructed a test set from the held-out patients Notably, the patients in the test sets had no overlap with those in the training set. The extracted patient-specific features from $\mathcal{M}(\cdot)$ effectively distinguished whether two WSIs belonged to the same patient, achieving **AUC** $=$ **0.9594**, **ACC** $=$ **0.8853**, and **F1-score** $=$ **0.8844**. These results demonstrate that the model successfully captures patient-specific information from WSIs.

Next, we investigate the extent to which patient-specific information is retained in slide embeddings under different settings: (1) the patient-specific feature extraction model, (2) the baseline MIL model, (3) the baseline with our method but uses direct slide-level concatenation, and (4) the baseline with our method. We selected 5 patients, each with multiple WSIs, and extracted their slide-level embeddings under these conditions. We then applied t-SNE for dimension reduction and visualization, examining the separability of WSIs based on patient identity. Embeddings from the patient-specific feature extraction model show clear clustering by patient identity. Baseline MIL model embeddings exhibit strong patient-specific biases, with WSIs from the same patient clustering together. Incorporating our framework significantly reduces this clustering effect, demonstrating its effectiveness in mitigating patient-specific biases in slide embeddings.

Analysis of Attention Regions. We analyze whether our framework improves the accuracy of attention regions. As shown in Fig. 4, the green outlines indicate regions highlighted by pathologists as representative of the key features. In the baseline model, high-attention patches mostly fall outside the annotated regions, whereas with our method, they align within these regions. This indicates that our approach enables the model to learn diagnostically relevant features.

4 Conclusions

In this study, we aimed to mitigate patient-specific biases in STS classification from WSIs. These biases introduce spurious correlations that mislead models and hinder generalization. To address this, we proposed a novel MIL framework to reduce the influence of patient-specific features. Extensive experiments on two STS datasets demonstrated the effectiveness of our approach, consistently improving classification performance. By suppressing patient-specific biases, our method enhances model generalization and reliability, contributing to a more accurate and clinically meaningful STS classification. The code, trained models, and testing pipeline will be available, enabling direct application in clinical settings to assist doctors in decision-making.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. 62371409) and Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005)

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

References

1. Chen, K., Sun, S., Zhao, J.: Camil: Causal multiple instance learning for whole slide image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1120–1128 (2024)
2. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
3. Crombé, A., Roulleau-Dugage, M., Italiano, A.: The diagnosis, classification, and treatment of sarcoma in this era of artificial intelligence and immunotherapy. Cancer Communications **42**(12), 1288–1313 (2022)
4. Foersch, S., Eckstein, M., Wagner, D.C., Gach, F., Woerl, A.C., Geiger, J., Glasner, C., Schelbert, S., Schulz, S., Porubsky, S., et al.: Deep learning for diagnosis and survival prediction in soft tissue sarcoma. Annals of Oncology **32**(9), 1178–1187 (2021)
5. Frankel, A.O., Lathara, M., Shaw, C.Y., Wogmon, O., Jackson, J.M., Clark, M.M., Eshraghi, N., Keenen, S.E., Woods, A.D., Purohit, R., et al.: Machine learning for rhabdomyosarcoma histopathology. Modern Pathology **35**(9), 1193–1203 (2022)
6. Hagi, T., Nakamura, T., Yuasa, H., Uchida, K., Asanuma, K., Sudo, A., Wakabayahsi, T., Morita, K.: Prediction of prognosis using artificial intelligence-based histopathological image analysis in patients with soft tissue sarcomas. Cancer Medicine **13**(10), e7252 (2024)
7. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)

8. 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)
9. Lin, W., Zhuang, Z., Yu, L., Wang, L.: Boosting multiple instance learning models for whole slide image classification: A model-agnostic framework based on counterfactual inference. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3477–3485 (2024)
10. 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)
11. Milewski, D., Jung, H., Brown, G.T., Liu, Y., Somerville, B., Lisle, C., Ladanyi, M., Rudzinski, E.R., Choo-Wosoba, H., Barkauskas, D.A., et al.: Predicting molecular subtype and survival of rhabdomyosarcoma patients using deep learning of h&e images: a report from the children's oncology group. Clinical Cancer Research **29**(2), 364–378 (2023)
12. Morisi, A., Rai, T., Bacon, N.J., Thomas, S.A., Bober, M., Wells, K., Dark, M.J., Aboellail, T., Bacci, B., La Ragione, R.M.: Detection of necrosis in digitised whole-slide images for better grading of canine soft-tissue sarcomas using machine-learning. Veterinary sciences **10**(1), 45 (2023)
13. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. 2016. Internet resource (2016)
14. 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)
15. Singer, S., Demetri, G.D., Baldini, E.H., Fletcher, C.D.: Management of soft-tissue sarcomas: an overview and update. The lancet oncology **1**(2), 75–85 (2000)
16. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
17. Vaidya, A., Chen, R.J., Williamson, D.F., Song, A.H., Jaume, G., Yang, Y., Hartvigsen, T., Dyer, E.C., Lu, M.Y., Lipkova, J., et al.: Demographic bias in misdiagnosis by computational pathology models. Nature Medicine **30**(4), 1174–1190 (2024)
18. Yeung, M.C., Cheng, I.S.: Artificial intelligence significantly improves the diagnostic accuracy of deep myxoid soft tissue lesions in histology. Scientific Reports **12**(1), 6965 (2022)
19. Zehra, T., Anjum, S., Mahmood, T., Shams, M., Sultan, B.A., Ahmad, Z., Alsubaie, N., Ahmed, S.: A novel deep learning-based mitosis recognition approach and dataset for uterine leiomyosarcoma histopathology. Cancers **14**(15), 3785 (2022)
20. Zhang, X., Wang, S., Rudzinski, E.R., Agarwal, S., Rong, R., Barkauskas, D.A., Daescu, O., Cline, L.F., Venkatramani, R., Xie, Y., et al.: Deep learning of rhabdomyosarcoma pathology images for classification and survival outcome prediction. The American Journal of Pathology **192**(6), 917–925 (2022)