Buffer-MIL: Robust Multi-instance Learning with a Buffer-based Approach

Gianpaolo Bontempo[1,2], Luca Lumetti[1], Angelo Porrello[1], Federico Bolelli[1], Simone Calderara[1], and Elisa Ficarra[1]

[1] University of Modena and Reggio Emilia, Italy
{name.surname}@unimore.it
[2] University of Pisa, Italy
{name.surname}@phd.unipi.it

Abstract. Histopathological image analysis is a critical area of research with the potential to aid pathologists in faster and more accurate diagnoses. However, Whole-Slide Images (WSIs) present challenges for deep learning frameworks due to their large size and lack of pixel-level annotations. Multi-Instance Learning (MIL) is a popular approach that can be employed for handling WSIs, treating each slide as a *bag* composed of multiple patches or *instances*. In this work we propose Buffer-MIL, which aims at tackling the covariate shift and class imbalance characterizing most of the existing histopathological datasets. With this goal, a buffer containing the most representative instances of each disease-positive slide of the training set is incorporated into our model. An attention mechanism is then used to compare all the instances against the buffer, to find the most critical ones in a given slide. We evaluate Buffer-MIL on two publicly available WSI datasets, Camelyon16 and TCGA lung cancer, outperforming current state-of-the-art models by 2.2% of accuracy on Camelyon16.

Keywords: Multi-instance Learning · Weakly Supervised Learning · Whole Slide Images

1 Introduction

The histopathological image analysis is a research area with a wide interest as it helps pathologists to carry out accurate diagnosis [12], especially when combined with genomic features [7,14,19]. The most common way to acquire glass slides is by employing Whole-Slide Image (WSI) scanners, which can produce digital high-resolution images [18]. Such resolutions are usually prohibitive for standard deep learning frameworks, and generating pixel-level accurate annotations represent a time-consuming and labor-intensive task. As a consequence, different strategies must be employed to perform automatic WSIs analysis and support clinicians in the daily practice. One of the most common approaches in literature follows the Multi-Instance Learning (MIL) paradigm, where from each slide (bag) multiple unlabelled patches (instances) are extracted. These patches have

a much smaller size w.r.t. the original image and can be directly fed into a deep learning network to obtain a positive or negative prediction (*e.g.,* tumor/not tumor). Once all the patch predictions are obtained, they must be aggregated to provide the final outcome for the entire slide. Indeed, bags can be perceived as a mosaic of interrelated concepts that are comprehensible only when viewed in their entirety [23].

Unfortunately, when dealing with positive bags, we also face the problem of class imbalance, as positive instances usually represent a low percentage of the entire set. Without correct precautions, the model will tend to overfit, and it might misclassify positive instances, leading to a wrong bag-level prediction. A second problem, named covariate shift, occurs when the distribution of instances within positive and negative bags differs between train and test data. This difference can force the model to focus on instances that are not actually related to the correct label [26]. This becomes crucial when dealing with *one-vs-all* cross attention paradigm [13], since the most critical instance drive the attention of all the others. Conversely, the *all-vs-all* attention (*e.g.,* self-attention in transformers) [4,8,11] approach can suffer from high-class imbalance, with instances that are often heterogeneous and noisy, making many comparisons irrelevant and even potentially derailing the final decision.

Motivated by the aforementioned challenges, this work proposes Buffer-MIL to address both class imbalance and covariate shift. To achieve this, our approach incorporates a *buffer-vs-all* strategy that makes use of a buffer to keep track of the most important instances seen during all the training process. This buffer is updated at run-time by selecting the top-k most critical instances of each positive slide in the training set. An attention mechanism is used to compare all the instances against the buffer, enabling the selection of the most critical ones to be incorporated into the learning process. This way, since the morphology of critical instances is more robust to covariate shift, we can leverage their stability to enhance the generalization performance of the model. We evaluate our approach on two publicly available WSI datasets, Camelyon16 and TCGA lung cancer, which demonstrate the effectiveness of the proposed approach. Specifically, Buffer-MIL outperforms the current state-of-the-art models by 2.2% in terms of accuracy and by 2.0% in terms of AUC on a single-scale setting.

Overall, our proposed Buffer-MIL approach provides an effective solution to address both class imbalance and covariate shift in classification tasks by leveraging a buffer containing the most critical instances, which allows for improved model performance. The source-code is available at `https://github.com/aimagelab/mil4wsi`.

2 Related Work

Multi-instance learning is a popular and well established type of supervised learning, whose application to the classification of WSIs is well known [3,11,13]. In this section, recent proposals about the application of MIL to WSIs are summarized, and the covariate shift problem is introduced.

2.1 Multi-instance Learning for WSI Analysis

Initially proposed for drug activity prediction [9], the multi-instance learning paradigm gained prominence in the world of histological whole-slide image analysis. Although initially employed as a simple instance classifier, recent studies introduce an attention mechanism to extract bag representations [2,6,13,16,17,20]. Among them, DS-MIL [13] is based on a dual-stream architecture. Patches are extracted from each considered magnification (5× and 20× in their study) of the WSIs and used (separately) for self-supervised contrastive learning. Patch embeddings extracted at different resolutions are later concatenated to train the MIL aggregator, which assigns an importance (or criticality) score to each instance. The most critical patch is then selected and compared to all the others (*one-vs-all*). Such comparison is based on a distance measure that recalls an attention mechanism, but it has a substantial difference as two queries are compared instead of using the classical *key and query* approach. All the distances are then aggregated into the final bag-level prediction. Differently, Ilse *et al.* [11] propose a MIL framework (AB-MIL) where the final aggregation function is based on a weighted average. The weights assigned to each instance are computed by a gated attention mechanism. The aim of this method is to find key instances in a fully differentiable and adaptable way, by comparing instances within a bag in an *all-vs-all* fashion.

2.2 Covariate Shift

Covariate shift refers to a marginal training distribution $P_{train}(X)$ that differs from the test one $P_{test}(X)$, maintaining stable the conditional distribution $P(y|X)$ [10,21]. In other words, we have a distribution shift when the training and the test set are not independent and identically distributed. This characteristic lead a neural network to learn features that are not correlated with the correct label. To mitigate these effects a widely used approach is importance weighting, which involves assigning a weight to each training instance x. This weight, denoted as $w(x)$, is calculated as the ratio of the marginal probabilities of the instance in the test and train sets, *i.e.,* $w(x) = P_{test}(X)/P_{train}(X)$. The weight-based approach aims at reducing the discrepancy between the train and test marginals improving the generalization performance of the model [22].

As observed in Stable-MIL [26], in covariate shift settings the meaning and characteristics of noisy instances may change due to the distribution differences between train and test sets. However, critical instances, characterized by their morphology or inherent properties, tend to remain stable and consistent regardless of the covariate shift. In other words, they exhibit robustness to the distribution changes and their predictive behavior remains reliable. Therefore, by focusing on instances that are less affected by the covariate shift, we can improve model stability to also enhance the generalization performance. In our approach, we adopt an attention module to automatically identify these critical instances and store them in a buffer for further analysis and integration into the model. Such buffer is then compared against all the instances of a bag to find patches with the highest contribution.

3 Model

3.1 Notation

Firstly, the notation that will be later used in this paper is introduced to better define the concepts described. With X, X^+, and X^- are denoted generic, positive, and negative bag respectively. Instead, with x we refer to a single instance extracted from a bag.

3.2 Critical Instances

The proposed multi-instance learning framework relies on the concept of critical instances, which play a fundamental role in determining the bag label. Formally, we define x as critical if it satisfies the following two conditions:

- x belongs to a positive bag X^+;
- adding x to a negative bag X^- would change the bag's label from negative to positive, that is, $\phi(X^- \cup \{x\}) = 1$, where ϕ is the function that maps a bag to its label.

The first condition ensures that the critical instance is informative about the positive class, while the second guarantees that the instance is not present in any negative bag that should have a positive label. Thus, critical instances are those that provide evidence for the positive class and cannot be easily explained away as noise. Intuitively, critical instances, x_{crit}, contain the most important information for bag classification. On the other hand, non-critical instances, x_{noisy}, may still contribute to the overall decision but their presence or absence does not have a significant impact on the outcome.

Assumption 1 *Critical instances exhibit similar patterns, unlike x_{noisy}. So, given a feature extractor f pretrained via a self-supervised paradigm, the similarity distance $d(\cdot,\cdot)$ across critical instances is lower than the one with other non-critical instances:*

$$d(f(x_{crit}), f(x_{crit})) < d(f(x_{crit}), f(x_{noisy})) \tag{1}$$

Starting from this assumption, our model builds a buffer containing most critical instances within each positive bag X^+, which is later used to measure how other instances are relevant. Since built over the entire training set, the buffer usage provides a wider knowledge about what is really important w.r.t. using a single instance, as done by DS-MIL.

3.3 Critical Buffer

To rank instances based on their importance within each slide, a standard attention-based DS-MIL [13] is employed. In particular, given a patch x, its embedding is computed as $h = f(x)$, where the function $f(\cdot)$ is obtained from

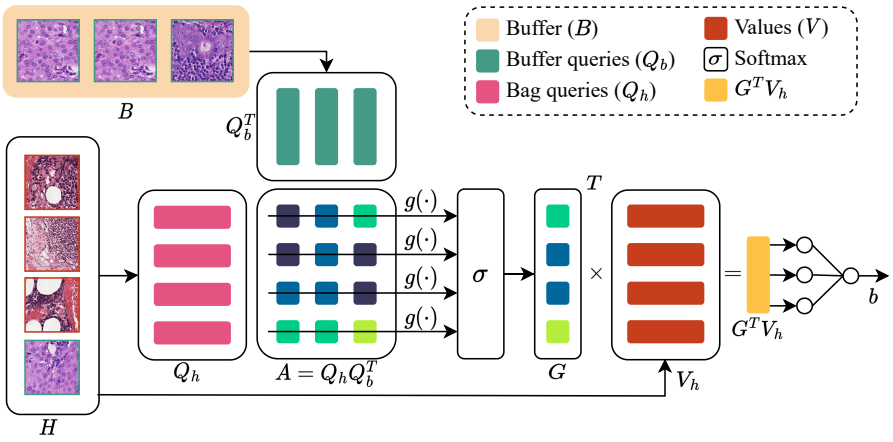

Fig. 1: Visual representation of the proposed model. In particular, given the buffer B and the input slide H, the attention matrix A is computed. The $g(\cdot)$ function is used to select the most informative elements from the matrix into G.

a self-supervised approach. A patch-level classifier $cls_{patch}(\cdot)$ is used to find the index of the most critical patch as:

$$\text{crit} = \text{argmax}(cls_{patch}(f(x))) = \text{argmax}\{W_p f(x_0), ..., W_p f(x_n)\} \tag{2}$$

where W_p is a weight vector.
The second step is to aggregate instance embeddings into a single bag embedding. This is performed by computing a linear projection of each embedding into a query q_i and a value v_i, using two weight matrices W_q and W_v:

$$\text{q}_\text{i} = W_q h_i, \quad \text{v}_\text{i} = W_v h_i \tag{3}$$

Next, the query relative to the most critical instance, q_{crit}, is obtained and compared to all other queries q_i (including itself) using a distance measure $U(\cdot, \cdot)$ defined as:

$$U(h_i, h_{\text{crit}}) = \frac{\exp(\langle q_i, q_{\text{crit}} \rangle)}{\sum_{k=0}^{N-1} \exp(\langle q_k, q_{\text{crit}} \rangle)} \tag{4}$$

Finally, the bag score is given by:

$$c_b(B) = W_b \sum_{i=0}^{N-1} U(h_i, h_{\text{crit}}) v_i \tag{5}$$

where W_b is again a weight vector. The bag score is used to select all the positive bags and extract the top-k instances within each of them. The ranking is given by the score $U(h_i, h_{\text{crit}})$. The buffer is build by training the aforementioned model; at the end of the process it contains the most critical instances of each bag, providing a more stable criticality representation.

The selection of the N most important patches from each slide (N/Slide) is repeated every $freq$ epochs, since the network should learn to assign a better score to bags and instances, better understanding what should actually be considered as critical.

3.4 Bag Embedding through the Critical Buffer

Fig. 1 illustrates how the buffer B is introduced in the attention mechanism. Given the current bag $H = \{h_1, ..., h_i, ..., h_N\}$, composed of N instances, and the buffer $B = \{b_1, ..., b_i, ..., b_M\}$, composed of M critical instances belonging from different slides, a new bag embedding can be computed. First, the weight matrix W_q trained in the previously described steps is used to perform a linear projection of all the instances h_i and all the instances within the buffer b_i, obtaining q_{h_i} and q_{b_i} respectively. An attention matrix A is then built, where $A_{i,j} = \langle q_{h_i}, q_{b_i} \rangle$. This can also be seen as a matrix multiplication, once defined $Q_h \in \mathcal{M}^{N \times K}$ as the row-wise concatenation of every q_{h_i} and $Q_b \in \mathcal{M}^{M \times K}$ as the row-wise concatenation of q_{b_i}, considering K the latent space size where each instance get projected, the attention matrix $A \in \mathcal{M}^{N \times M}$ can be written as follow:

$$A = Q_h Q_b^T \tag{6}$$

As only a single attention score is required for each of the bag instances h_i, an aggregation function $g(\cdot)$ on each row of A must be used to obtain a new matrix $G \in \mathcal{M}^{N \times 1}$ as $G_i = g(\{A_{i,j} : \forall j \in [1, M]\})$.

All the instances h_i are also projected into values v_{h_i} of size L using the W_v weight matrix of the previous step, obtaining $V_h \in \mathcal{M}^{N \times L}$. Finally, the bag embedding is computed as:

$$b = W_b G^T V_h \tag{7}$$

with $W_b \in \mathcal{M}^{1 \times L}$ representing the weight matrix that computes the final bag embedding. In this paper, two different function $g(\cdot)$ are proposed:

- **mean**: the attention scores are computed considering the entire buffer, under the assumption that it is composed of critical instances only. In particular $G_i = \text{mean}\{A_{ij} : \forall j \in [1, M]\}$;
- **max**: considering that the buffer may also contain noisy labels, using a max-pooling operation allows to select only the most representative instances. Specifically, $G_i = \max\{A_{ij} : \forall j \in [1, M]\}$

4 Experimental Settings and Results

4.1 Pre-processing

Each slide has been cropped using the CLAM framework [15], a state-of-the-art tool for selecting tissue patches and removing the WSI background. In particular, each slide has been processed at thumbnail level through a combination of Otsu thresholding [25] and connected components analysis [1], to obtain the

tissue contours. After that, each 256×256 patch within the selected contours is extracted without overlapping at $20\times$ scale resolution ($5\times$ and $20\times$ in the multi-scale setting).

Finally, instance embeddings are obtained through a ViT model trained in a self-supervised fashion by means of the DINO paradigm [5]. The training is performed separately on each dataset/resolution. The model has been trained for a week with two NVIDIA GeForce GTX 2080 Ti GPUs using the default parameters proposed by the authors.

4.2 Metrics

The evaluation metrics considered are the Area Under the Curve (AUC) and the accuracy. As the name suggests, the AUC measures the area under the ROC curve, representing the relationship between the true positive rate, $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, and the false positive rate, $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$, for any possible threshold. Once the best threshold for the ROC curve is found, we measure the accuracy as the quantity of TP over the entire test set. Each experiment has been executed with 3 different seeds, reporting the average and the standard deviation .

4.3 Datasets

The proposed method has been extensively tested over two different datasets: Camelyon16 and TCGA Lung. The former has been created with the purpose of automatic detection of metastases in Hematoxylin and Eosin (H&E) stained whole-slide images of lymph node sections, as part of the homonymous challenge held at the International Symposium on Biomedical Imaging (ISBI) in 2016 [2]. The dataset comprises a total of 398 WSIs, out of which 128 are designated as "official test set". The images were acquired through two slide scanners, namely RUMC and UMCU, respectively equipped with $20\times$ and $40\times$ objective lenses. The specimen-level pixel sizes are comparable, *i.e.,* $0.243\mu m \times 0.243\mu$m for RUMC and 0.226μm $\times$ 0.226μm for UMCU. Official training and test set have been employed for our experiments.

The second dataset, publicly available on the GDC Data Transfer Portal, comprises two sub-types of cancer: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC), counting 541 and 513 WSIs respectively. In this case, the task is the classification of LUAD vs LUSC. To provide a fair comparison with Li *et al.* [13], we employ the same split between train and test set and remove ten corrupted slides as suggested in the original publication.

4.4 Results

Tab. 1 compares the proposed Buffer-MIL with state-of-the-art approaches: two MIL models with simple aggregators like mean-pooling and max-pooling, Attention-based MIL (AB-MIL) [11], DS-MIL, and its multi-scale version [13]. We also extend the buffer-based approach to consider multiple resolutions.

Table 1: Performance comparison on Camelyon16 and TCGA Lung dataset. The "†" identifies multi-scale approaches. Buffer aggregation is based on mean in these experiments.

	Camelyon16		**TCGA Lung**	
Model	**Accuracy**	**AUC**	**Accuracy**	**AUC**
mean-pooling	0.723 ± 0.004	0.672 ± 0.010	0.823 ± 0.002	0.905 ± 0.001
max-pooling	0.893 ± 0.015	0.899 ± 0.007	0.851 ± 0.008	0.909 ± 0.002
AB-MIL	0.724 ± 0.015	0.744 ± 0.016	0.864 ± 0.009	0.933 ± 0.004
DS-MIL	0.915 ± 0.013	0.952 ± 0.005	0.888 ± 0.005	$\mathbf{0.951 \pm 0.002}$
Buffer-MIL	$\mathbf{0.935 \pm 0.012}$	$\mathbf{0.971 \pm 0.005}$	$\mathbf{0.891 \pm 0.008}$	0.950 ± 0.002
DS-MIL†	0.909 ± 0.020	0.955 ± 0.010	$\mathbf{0.913 \pm 0.005}$	$\mathbf{0.966 \pm 0.002}$
Buffer-MIL†	$\mathbf{0.940 \pm 0.008}$	$\mathbf{0.969 \pm 0.005}$	0.897 ± 0.020	0.956 ± 0.010

Table 2: Comparison between the usage of max and mean aggregation (Agg.) by setting the buffer update frequency to 10.

Agg.	**N/slide**	**Accuracy**	**AUC**
Mean	1	0.934 ± 0.012	0.970 ± 0.006
	2	0.932 ± 0.012	0.968 ± 0.006
	10	$\mathbf{0.935 \pm 0.012}$	$\mathbf{0.971 \pm 0.005}$
Max	1	0.925 ± 0.012	0.966 ± 0.004
	2	0.927 ± 0.020	0.967 ± 0.005
	10	0.930 ± 0.021	0.967 ± 0.003

From a single scale perspective, using the buffer improves the baseline by an average of 2.2% in accuracy and 2.0% in AUC on the Camelyon16 and 0.3% in accuracy for the TCGA Lung dataset. Employing multiple resolutions generally provide better performances: on Camelyon16 the buffer improves the baseline by an average of 3.4% in accuracy and 1.5% in AUC.

4.5 Model Analysis

Our experiments provide evidence that Buffer-MIL is effective at tackling covariate shift, as demonstrated by the higher performance improvement obtained on Camelyon16 compared to TCGA Lung (Tab. 1). Given the smaller size of Camelyon16, overfitting can become a critical issue, slightly attenuated by the multi-scale approach.

Aggregation Function. Two different aggregation functions have been studied and presented in Tab 2. Experimental results reveal that producing the final attention scores by averaging critical representations in the buffer outperforms the use of a max operator.

Table 3: Contribution of buffer update frequency (Freq.) when using mean-based aggregation.

Freq.	N/slide	Accuracy	AUC
1	10	0.919 ± 0.012	0.963 ± 0.004
2	10	0.917 ± 0.009	0.967 ± 0.001
10	10	$\mathbf{0.935 \pm 0.012}$	$\mathbf{0.971 \pm 0.005}$

One possible explanation is that selecting only the most representative disease-positive buffer instances produces a final representation that is not aligned with all the bags. This approach may not capture the diversity of the disease-positive instances and may lead to sub-optimal performance. In contrast, the mean operator takes into account all the critical instances, which allows for a stronger consensus. This approach is better at capturing the diversity of disease-positive instances and is less likely to overfit specific patches. Furthermore, the mean operator is less sensitive to outliers and noise that may be contained in the buffer.

Buffer Update Frequency. This hyperparameter regulates the interval (measured in epochs) between each buffer update. In Tab. 3, we also investigate the impact of buffer update frequency, which is found to be an important parameter for both max and mean operators.

Our analysis suggests that updating the buffer fewer times generally leads to better performances, as it allows for a better selection of the most representative disease-positive instances across the entire training set. Updating the buffer with an higher frequency prevents its consolidation, and may cause it to be filled with noisy or irrelevant information. Instead, updating the buffer less frequently increases the time interval between buffer creations, causing it to become outdated and failing to capture the most relevant instances. Setting an appropriate interval is required by the model to learn and generalize from the initial training data before incorporating new information into the buffer. In other words, the model can better consolidate the knowledge from the initial training data, and, consequently, perform a better selection of new instances. It is essential to find the right trade-off.

Buffer Size. The buffer is built considering the N most critical instances from each slide. As illustrated in Tab. 4, our analysis demonstrates that the impact of buffer size is less significant w.r.t. buffer update frequency. Our experiments also suggest that increasing the buffer size does not always lead to improved performance.

One possible explanation is that when the buffer frequency update is low, increasing the buffer size may include more irrelevant or noisy instances, which could negatively impact the model performance. In this scenario, selecting a larger number of instances per slide could cause the buffer to become more "diluted" with irrelevant instances. As a result, the model may not be able to

Table 4: Buffer size contribution at different update frequencies when using mean-based aggregation.

Freq	N/slide	Accuracy	AUC
1	1	0.922 ± 0.008	0.962 ± 0.002
	2	$\mathbf{0.925 \pm 0.012}$	$\mathbf{0.961 \pm 0.003}$
	10	0.919 ± 0.012	0.963 ± 0.004
10	1	0.934 ± 0.012	0.970 ± 0.006
	2	0.932 ± 0.012	0.968 ± 0.006
	10	$\mathbf{0.935 \pm 0.012}$	$\mathbf{0.971 \pm 0.005}$

Table 5: Comparison with random sampling when using mean-based aggregation and a frequency update of 10.

	Our Method		Reservoir Sampling	
N/slide	Accuracy	AUC	Accuracy	AUC
1	0.934 ± 0.012	0.970 ± 0.006	0.922 ± 0.014	0.962 ± 0.003
2	0.932 ± 0.012	0.968 ± 0.006	0.922 ± 0.008	0.963 ± 0.004
10	$\mathbf{0.935 \pm 0.012}$	$\mathbf{0.971 \pm 0.005}$	0.925 ± 0.012	0.964 ± 0.004

properly consolidate and learn from the most critical instances, leading to a decrease in performance.

On the other hand, when the buffer update frequency is high, the buffer can better capture the most critical disease-positive instances, even if the buffer size is small. In this case, the mean operator typically works better on bigger buffers, but small buffer sizes can still perform comparably well. Selecting the optimal buffer size depends on the specific dataset and task, as well as the buffer update frequency.

Sampling Selection. To provide evidence that selecting proper patches matter, in Tab. 5 we show a comparison between our proposed method, and the reservoir sampling strategy [24], which is a random-based selection technique. The results demonstrate that our approach outperforms the random selection strategy regardless of the parameters used.

5 Conclusion

In conclusion, our analysis demonstrates that Buffer-MIL is an effective approach for addressing the problem of covariate shift when multi-instance learning is applied to the histopathological context. In particular, the results suggest that performing an appropriate buffer selection approach and identifying the correct interval for updating the buffer are critical to achieve optimal performance.

Further research is needed to investigate how relevant buffers are in more difficult and diverse tasks such as survival prediction. In that case, tissue morphology is not directly connected to the patient outcome and a better storage strategy (*e.g.,* multiple buffers per concept) would be probably needed.

Acknowledgements This project has received funding from DECIDER, the European Union's Horizon 2020 research and innovation programme under GA No. 965193, and from the Department of Engineering "Enzo Ferrari" of the University of Modena through the FARD-2022 (Fondo di Ateneo per la Ricerca 2022).

References

1. Allegretti, S., Bolelli, F., Cancilla, M., Pollastri, F., Canalini, L., Grana, C.: How does Connected Components Labeling with Decision Trees perform on GPUs? In: Computer Analysis of Images and Patterns. pp. 39–51. Springer (2019) 6
2. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., 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) 3, 7
3. Bontempo, G., Porrello, A., Bolelli, F., Calderara, S., Ficarra, E.: DAS-MIL: Distilling Across Scales for MIL Classification of Histological WSIs. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (2023) 2
4. Bruno, P., Amoroso, R., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: Investigating Bidimensional Downsampling in Vision Transformer Models. In: Image Analysis and Processing – ICIAP 2022. pp. 287–299. Springer (2022) 2
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021) 7
6. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16144–16155 (2022) 3
7. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F.: Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. IEEE Transactions on Medical Imaging **41**(4), 757–770 (2020) 1
8. Cornia, M., Baraldi, L., Cucchiara, R.: Explaining Transformer-based Image Captioning Models: An Empirical Analysis. AI Communications **35**(2), 111–129 (2022) 2
9. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence **89**(1), 31–71 (1997) 3
10. Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., Smola, A.: Correcting Sample Selection Bias by Unlabeled Data. Advances in Neural Information Processing Systems 19 (NIPS) **19** (2006) 3
11. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: International Conference on Machine Learning. vol. 80, pp. 2127–2136. PMLR (Jul 2018) 2, 3, 7

12. Kumar, N., Gupta, R., Gupta, S.: Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions. Journal of Digital Imaging **33**(4), 1034–1040 (2020) 1
13. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14318–14328 (2021) 2, 3, 4, 7
14. Lovino, M., Bontempo, G., Cirrincione, G., Ficarra, E.: Multi-omics Classification on Kidney Samples Exploiting Uncertainty-Aware Models. In: Intelligent Computing Theories and Application. pp. 32–42. Springer (2020) 1
15. 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) 6
16. Maksoud, S., Zhao, K., Hobson, P., Jennings, A., Lovell, B.C.: SOS: Selective Objective Switch for Rapid Immunofluorescence Whole Slide Image Classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3862–3871 (2020) 3
17. Panariello, A., Porrello, A., Calderara, S., Cucchiara, R.: Consistency-Based Self-supervised Learning for Temporal Anomaly Localization. In: Computer Vision – ECCV 2022 Workshops. pp. 338—-349 (2022) 3
18. Ponzio, F., Urgese, G., Ficarra, E., Di Cataldo, S.: Dealing with Lack of Training Data for Convolutional Neural Networks: The Case of Digital Pathology. Electronics **8**(3) (2019) 1
19. Roberti, I., Lovino, M., Di Cataldo, S., Ficarra, E., Urgese, G.: Exploiting Gene Expression Profiles for the Automated Prediction of Connectivity between Brain Regions. International Journal of Molecular Sciences **20**(8), 2035 (2019) 1
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 (NeurIPS) **34**, 2136–2147 (2021) 3
21. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference **90**(2), 227–244 (2000) 3
22. Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., Kawanabe, M.: Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. Advances in Neural Information Processing Systems 20 (NIPS) **20** (2007) 3
23. Tu, M., Huang, J., He, X., Zhou, B.: Multiple instance learning with graph neural networks. In: ICML Workshop on Learning and Reasoning with Graph-Structured Representations (2019) 2
24. Vitter, J.S.: Random Sampling with a Reservoir. ACM Transactions on Mathematical Software **11**(1), 37–57 (1985) 10
25. Zhang, J., Hu, J.: Image Segmentation Based on 2D Otsu Method with Histogram Analysis. In: International Conference on Computer Science and Software Engineering. vol. 6, pp. 105–108. IEEE (2008) 6
26. Zhang, W., Li, J., Liu, L.: Robust Multi-Instance Learning with Stable Instances. In: ECAI 2020: 24th European Conference on Artificial Intelligence (2019) 2, 3