Multi-scale Attention-based Multiple Instance Learning for Breast Cancer Diagnosis

Mariana Mourão*, Jacinto C. Nascimento, Carlos Santiago, and Margarida Silveira

Institute for Systems and Robotics, Instituto Superior Técnico, Lisbon, Portugal
Corresponding author: marianamourao@tecnico.ulisboa.pt

Abstract. Multiple Instance Learning (MIL) is a powerful weakly supervised learning framework for high-resolution medical images, but its application in mammographic breast cancer (BC) diagnosis overlooks instance interactions and the multi-scale nature of BC lesions. In this work, we propose a novel Feature Pyramid Network (FPN)-MIL model for BC classification and detection in high-resolution mammograms, integrating **(1)** a FPN-based instance encoder that enables a multi-scale analysis across different receptive-field granularities while operating on single-scale input patches; **(2)** deep-supervised scale-specific instance aggregators that support conventional attention (AbMIL) or transformer-based (SetTrans) mechanisms; **(3)** an attention-based multi-scale aggregator that dynamically combines scale-specific features, improving robustness to lesion scale variability. Our experiments show that FPN-MIL is superior to conventional single- and multi-scale patch-based MIL models, with FPN-SetTrans outperforming baselines in calcification classification and detection while FPN-AbMIL performs best for mass classification. Code is available publicly at: https://github.com/marianamourao-37/Multi-scale-Attention-based-MIL.

Keywords: Mammography · Multiple instance learning (MIL) · Feature Pyramid Network (FPN) · Transformer

1 Introduction

Breast cancer (BC) is the most diagnosed cancer worldwide, with over 3 million new cases and 1 million related deaths estimated by 2040 [1]. Mammography is the gold standard for early BC detection, providing high-resolution imaging of suspicious lesions (e.g., masses and calcifications) [19, 29]. While deep learning (DL)-based computer-aided diagnosis (CAD) systems have shown promise in mammographic BC diagnosis (MBCD), they face key challenges: **(1)** full image-based DL models typically rely on downsampled images, compromising robust feature learning for small Regions-of-Interest (ROIs), besides their "black-box" nature limiting interpretability [3, 21, 26]; **(2)** ROI-based DL models improve interpretability and achieve state-of-the-art performances, but require labor-intensive annotations (such as bounding-boxes or patch annotations) [5, 21, 26].

Multiple Instance Learning (MIL) has emerged as a powerful weakly supervised learning (WSL) framework for high-resolution medical images, treating them as a bag of instances (e.g., patches or pixels) that are aggregated for image-level classification while relying only on weak image-level supervision [5]. Early instance-based MIL models [6, 10, 28] focused on instance-level learning but suffered from noisy instance labels due to the lack of direct supervision, degrading image classification and instance localization [11]. In contrast, embedded-based MIL models transform the MIL problem into a standard supervised learning task by computing a joint bag embedding from instance features, typically achieving improved performances [5, 11]. Most embedded-based MIL research focuses on histopathologic whole-slide images [7, 14, 16, 17, 27], whereas MBCD studies primarily address instance ambiguity through conventional attention-based MIL aggregators [2, 3, 22, 23], overlooking instance interactions and the multi-scale nature of BC lesions. Transformer-based MIL aggregators address the former, including more efficient formulations for the commonly large-size bags in CAD applications [13]. Existing multi-scale MIL models typically operate on multi-scale input patches [7, 8, 14, 17, 27], increasing computational cost and limiting lesion detection granularity [12]. Alternatively, pixel-based MIL models [10, 20, 28] enhance localization granularity by treating feature-map pixels as instances but often rely on downsampled input images, losing fine-grained details [12].

In this work, we propose a novel embedded-based FPN-MIL model to classify and localize BC in full-resolution mammograms. Our main contributions are: **(1)** a FPN-based instance encoder enabling multi-scale analysis across different receptive-field granularities while operating on single-scale input patches; **(2)** Deep-supervised scale-specific instance aggregators that leverage hierarchical features, supporting either attention-based (AbMIL) or transformer-based (SetTrans) mechanisms; **(3)** An attention-based multi-scale aggregator that dynamically combines scale-specific features for a unified analysis, enhancing robustness to lesion scale variability; **(4)** Experiments show that our FPN-MIL is superior to conventional single/multi-scale patch-based MIL models, with FPN-SetTrans outperforming all baselines in calcification classification and detection while FPN-AbMIL performs best for mass classification. To the best of our knowledge, the proposed FPN-MIL is the first embedded-based MIL model to address the multi-scale nature of lesions and instance interactions in MBCD.

2 Method

The proposed FPN-MIL model is illustrated in Figure 1. Similar to a typical MIL framework for MBCD, an input grayscale mammogram $I \in \mathbb{R}^{H \times W}$ is converted into a grid of patches $B = \{b_i\}_{i=1}^{N}$, where N is the number of extracted patches and each patch $b_i \in \mathbb{R}^{H_p \times W_p}$ has dimensions (H_p, W_p). Unlike conventional MIL models that consider patch-level instances directly, a novel **FPN-based instance encoder** is introduced to hierarchically extract fine-to-coarse instance feature vectors X^s from multi-scale feature maps at different pyramid levels $s \in \{1, ..., S\}$. Deep-supervised **scale-specific instance aggregators** leverage the hierarchical features to independently compute bag embeddings h^s and predictions P^s, while the attention-based **multi-scale aggregator** adaptively integrates information across scales for a unified analysis. The following subsections provide a more detailed description of the main modules.

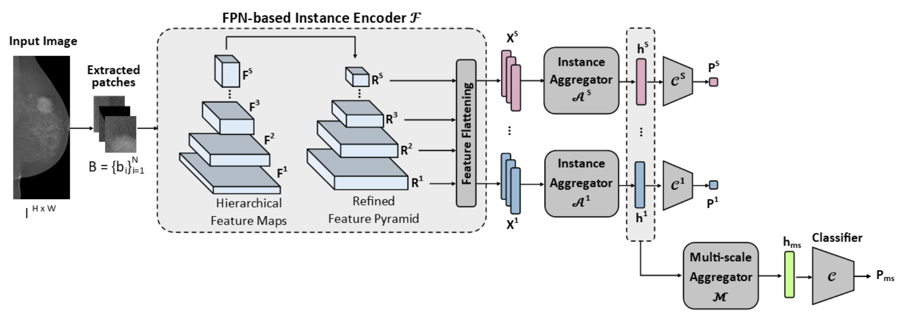

Fig. 1. Overview of the proposed FPN-MIL model. Deep-supervised instance aggregators leverage instance features X^s across pyramid levels s, computing bag embeddings h^s and predictions P^s. The attention-based multi-scale aggregator combines $\{h^s\}_{s=1}^S$ into a multi-scale bag embedding h_{ms} to produce the final prediction P_{ms}.

2.1 FPN-based Instance Encoder

The FPN-based Instance Encoder $\mathcal{F}$ consists of a shared hierarchical architecture that independently and identically processes each patch within a bag $b \in B$, generating instance feature vectors $X^s = \mathcal{F}(b) \in \mathbb{R}^{d_x \times \frac{H_p}{s} \times \frac{W_p}{s}}$ associated with feature-map pixels at different pyramid levels $s \in \{1, ..., S\}$. To address the semantic gap inherent in hierarchical backbones (e.g., CNNs), an FPN architecture is used to semantically refine the backbone's bottom-up feature maps $\{F^1, ..., F^S\}$ into a top-down feature pyramid $\{R^1, ..., R^S\}$. For simplicity, the original FPN architecture proposed by Lin et al. [15] was adopted, given by:

$$\begin{aligned} R^S &= \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{1\times1}(F^S)) \\ R^s &= \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{1\times1}(F^s) + \mathrm{Up}(R^{s+1})), s \in \{1, ..., S-1\}, \end{aligned} \tag{1}$$

where the 1×1 and 3×3 convolutional layers produce d_x-channel outputs, ensuring consistent feature dimension across the refined feature maps [15]. For the subsequent MIL framework, the 2D multi-scale feature maps $\{R^s\}_{s=1}^S$ are flattened to generate corresponding instance feature matrices $X^s = \{x_i^s\}_{i=1}^{n_s} \in \mathbb{R}^{n_s \times d_x}$, where the number of instances per scale is $n_s = N \times \frac{H_p}{s} \times \frac{W_p}{s}$.

2.2 Deep-supervised Scale-specific Instance Aggregators

Deep-supervised scale-specific instance aggregators are integrated to effectively leverage multi-scale information across pyramid levels, providing additional MIL supervision to enhance hierarchical feature learning as suggested by Wang et al. [25]. Each scale-specific instance aggregator $\mathcal{A}^s$ independently processes an instance feature matrix $X^s \in \mathbb{R}^{n_s \times d_x}$ into a corresponding bag embedding $h^s = \mathcal{A}^s(X^s) \in \mathbb{R}^d$, followed by a classification head $\mathcal{C}^s$ that computes the bag probability $P^s = \mathcal{C}^s(h^s) \in [0,1]$. In this work, we investigate two attention-based aggregators that can be decomposed into an encoder and a pooling stage. For ease of notation, the scale-specific superscript s will be omitted.

Attention-based MIL (AbMIL) In the pioneer work by Ilse et al. [11], the encoder stage employs an MLP to transform instance features $X \in \mathbb{R}^{n \times d_x}$ into lower-dimensional embeddings $Z = MLP(X) \in \mathbb{R}^{n \times d}$, with the encoded feature dimension d being an hyperparameter. The pooling stage consists of a learnable weighted-average operator:

$$h = \sum_{i=1}^{n} a_i z_i, \tag{2}$$

where attention weights a_i quantify each instance's contribution to the bag classification. These weights are computed through a specialized neural network with two fully connected layers parameterized by V, $U \in \mathbf{R}^{L \times d}$, followed by element-wise multiplication $\odot$ and a softmax normalization:

$$a_i = \frac{\exp\left\{w^\intercal(\tanh(Vz_i^\intercal) \odot \mathrm{sigm}(Uz_i^\intercal))\right\}}{\sum_{j=1}^{n} \exp\left\{w^\intercal(\tanh(Vz_j^\intercal) \odot \mathrm{sigm}(Uz_j^\intercal))\right\}}, \tag{3}$$

with the attention pooling dimension L being another hyperparameter. Instance-level attention scores $A = \{a_i\}_{i=1}^n$ are posteriorly used to produce interpretable heatmaps.

Set Transformer (SetTrans) It is a permutation-invariant transformer-based aggregator proposed by Lee et al. [13], with its basic operation being the Multi-head Attention Block (MAB):

$$\begin{aligned} \mathrm{MAB}(X,Y) &:= \mathrm{LN}(Z' + \mathrm{MLP}(Z')) \\ Z' &:= \mathrm{LN}(X + \mathrm{MHA}(X,Y,Y)), \end{aligned} \tag{4}$$

where LN denotes Layer Norm and MHA is the multi-head attention mechanism proposed in the original transformer [24]. For dealing with large-size bags, the permutation-equivariant Induced Set Attention Blocks (ISABs) are employed:

$$\mathrm{ISAB}_m(X) := \mathrm{MAB}(X, \mathrm{MAB}(I_m, X)), \tag{5}$$

relying on a set of m-trainable inducing points $I_m \in \mathbb{R}^{m \times d}$ to produce a contextually enriched encoded set $Z \in \mathbb{R}^{n \times d}$, notably reducing conventional computational complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(m.n)$. This encoder stage has some hyperparameters: the encoded feature dimension d; the number of inducing points m; the number of attention heads $n_{h,ISAB}$; the number of ISAB layers L_e, with

$L_e > 1$ capturing higher-order instance interactions. Since n varies across scales, a rule-based criterion is considered to set $m = 10 \times \log(n)$ that ensures $m \ll n$ for attaining computational efficiency across scales. Regarding the pooling stage, the permutation-invariant Pooling by Multi-head Attention (PMA) is employed:

$$\text{PMA}(Z) := \text{MAB}(S_e, Z), \tag{6}$$

relying on a learnable seed vector $S_e \in \mathbb{R}^{1\times d}$ as the query to aggregate the encoded bag feature matrix $Z \in \mathbb{R}^{n\times d}$ into a corresponding bag embedding $h = PMA(Z) \in \mathbb{R}^d$. The number of heads $n_{h,PMA}$ is an hyperparameter. Importantly, PMA also produces instance-level attention scores $A = \{a_i\}_{i=1}^n$ computed through the MHA mechanism.

2.3 Attention-based Multi-scale Aggregator

The attention-based multi-scale aggregator $\mathcal{M}$ computes a multi-scale bag embedding $h_{ms} = \mathcal{M}(H) \in \mathbb{R}^d$ by adaptively weighting the scale-specific bag embeddings $H = \{h^s\}_{s=1}^S$ using the AbMIL mechanism [11], with scale scores a^s given by:

$$a^s = \frac{\exp\left\{w^\top(\tanh(V{h^s}^\top) \odot \text{sigm}(U{h^s}^\top))\right\}}{\sum_{j=1}^S \exp\left\{w^\top(\tanh(V{h^j}^\top) \odot \text{sigm}(U{h^j}^\top))\right\}}. \tag{7}$$

Finally, a classification head $\mathcal{C}$ predicts the final bag probability $P_{ms} = \mathcal{C}(h_{ms}) \in [0, 1]$ which determines image-level classification.

3 Experiments

3.1 Dataset

The publicly available dataset VinDr-Mammo [18] was used to evaluate the performance of the proposed model, containing 5000 four-view exams with image-level assessment labels and annotated bounding-boxes for non-benign findings (e.g., mass, calcification). The original train-test split is used, with the training set further divided by a $80\% - 20\%$ stratified grouped split to obtain a validation set, used for monitoring the model's performance during training.

3.2 Experimental details

Data Pre-processing The pre-processed mammograms from the VinDr dataset provided by Ghosh et al. [9] were used. **Implementation Details** Patch-based MIL baselines (AbMIL [11] and SetTrans [13]) were implemented, operating on conventional 256×256 non-overlapping patches. In contrast, our FPN-MIL models process 512×512 non-overlapping patches for enabling a more comprehensive multi-scale analysis. Following prior deep MIL models that handle large-size bags [7, 14, 16, 17], we use a frozen pre-trained backbone for offline instance feature extraction. Specifically, the pre-trained Mammo-CLIP based on an EfficientNet-B2 (EN-B2) was chosen as a state-of-the-art Vision-Language foundational model for MBCD [9]. For patch-based MIL baselines, extracted instance features vectors have a dimensionality of $d_x = 352$. In our FPN-MIL models, the last two bottom-up feature maps were extracted offline and refined online into a top-down feature pyramid with a shared feature dimension of $d_x = 256$. To extend the multi-scale analysis to a larger scale, a stride-four downsampling was applied over the coarser feature maps similar to the approach of Lin et al. [15]. Regarding training configurations, we adopted a setup similar to Ghosh et al. [9] for the downstream classification task. Specifically, all MIL models were trained with a batch-size of 8 for 30 epochs using the AdamW optimizer with initial learning rate of 5e−5, a weight decay of 1e−4 and a cosine-annealing learning-rate scheduler. The official hyperparameters for AbMIL and SetTrans models were used, namely: $d = 256$; $L = 128$; $n_{h,ISAB} = 4$; $L_e = 2$. For model optimization, we applied a class-weighted binary cross-entropy loss across all scales, combining multi-scale and scale-specific losses. **Evaluation Metrics** The models are evaluated for classification and detection of masses and calcifications in the VinDr dataset. Binary image-level classification is reported using AUC-ROC, relying on ground-truth labels $\{\text{No } \langle E \rangle, \langle E \rangle\}$, where E denotes either a mass or calcification. Localization performance is evaluated in a post-hoc analysis of the multi-scale aggregated heatmaps, with mean Average Precision (mAP) being reported at an IoU threshold of 0.25. We also report mAP for lesions of different sizes: small (area $< 128^2$ pixels), medium ($128^2 <$ area $< 256^2$ pixels) and large (area $> 256^2$ pixels), respectively denoted as mAP_s, mAP_m and mAP_l. Following prior MIL works [2, 16], fine-grained heatmaps are generated during inference by defining a 75% overlap between extracted patches, where the attention scores in overlapped regions are accumulated and averaged. Instance-level attention scores are then re-scaled with min-max normalization and mapped to their corresponding spatial locations in the mammogram. The multi-scale aggregated heatmap is obtained by weighting scale-specific heatmaps according to the scale scores learned by the multi-scale aggregator. To generate predicted bounding-boxes, isolated high-attention regions from the heatmap are extracted by simultaneously thresholding pixel values above the 95% quantile of the heatmap's distribution [9] and a fixed threshold of 0.5 for further refinement.

4 Results and Discussion

4.1 Comparison with Baselines

Table 1 compares the proposed FPN-MIL models against baselines across different learning paradigms. For MIL models, SetTrans-based aggregators perform better for calcifications, possibly helping to recognize clusters of microcalcifications highly associated with malignancy [19] rather than treating them in isolation by establishing long-range instance interactions. Contrarily, masses are isolated volumes that seem to benefit from the localized nature of AbMIL-based aggregators that help preserve mass shape and structure. Notably, our FPN-MIL models significantly improve detection performance across lesion sizes compared with SSP-MIL baselines, being illustrated in Figure 2 the multi-scale aggregated heatmaps for our best-performing models. Specifically, FPN-SetTrans achieves the best performance in calcification classification and detection, while the FPN-AbMIL achieves the best mass classification but fails to surpass in mass detection compared to the FSOD and WSOD models. Given the greater variability in mass appearance and poorer contrast [4], our models struggles with accurate mass detection under the limited image-level supervision. While RetinaNet benefits from ground-truth bounding boxes for improved detection, Mammo-FActOR leverages an image-text alignment mechanism for sentence-level granularity [9] which proves particularly effective for mass detection possibly due to well-defined mass attributes (e.g., shape, size and margins) in the available radiology reports.

Table 1. Performance of the proposed FPN-MIL models compared with baselines across different learning paradigms: Fully Supervised Classification (FSC); Fully Supervised Object Detection (FSOD); Weakly Supervised Object Detection (WSOD); Single-scale Patch-based MIL (SSP-MIL). Detection performance is reported for all (mAP), small (mAP$_s$), medium (mAP$_m$) and large (mAP$_l$) lesions. Results for EN-B2, RetinaNet and Mammo-FActOR are reported from [9] under the linear probe setting.

Type	**Model**	**Calcification**					**Mass**				
		AUC	**mAP**	**mAP$_s$**	**mAP$_m$**	**mAP$_l$**	**AUC**	**mAP**	**mAP$_s$**	**mAP$_m$**	**mAP$_l$**
FSC	EN-B2 [9]	92.0	-	-	-	-	73.0	-	-	-	-
FSOD	RetinaNet [9]	-	17.0	-	-	-	-	37.0	-	-	-
WSOD	Mammo-FActOR [9]	-	20.0	-	-	-	-	**38.0**	-	-	-
SSP-MIL	AbMIL [11]	90.5	15.9	0.0	26.6	52.1	75.8	14.7	0.0	18.8	61.0
	SetTrans [13]	88.9	18.4	0.1	29.4	57.6	73.2	5.8	0.0	9.1	22.0
FPN-MIL	(Our) FPN-AbMIL	93.5	32.0	9.1	34.8	57.5	**79.2**	28.2	**4.7**	**32.1**	66.2
	(Our) FPN-SetTrans	**94.2**	**37.4**	**18.8**	**39.5**	**62.2**	77.4	24.3	3.0	28.0	**73.2**

4.2 Ablation Studies

The following ablation studies are conducted on the best-performing models (i.e., FPN-SetTrans for calcifications and FPN-AbMIL for masses), with results summarized in Table 2. **(1) Effect of FPN-based Instance Encoder:** We compare our FPN-based instance encoder against the conventional multi-scale patch (MSP) encoders while keeping the rest of the model preserved. Following our three-scale model design, we consider patch-sizes of 128, 256 and 384. The obtained results demonstrate the superiority of our FPN-based instance encoder in the classification and detection of both lesion types, particularly boosting small lesion detection given its improved receptive-field granularity over patch-level encoders. **(2) Effect of Multi-Scale Aggregator:** We also analyze the impact of different multi-scale aggregators in our FPN-MIL models. Removing the multi-scale aggregator (w/o MS-Aggr) results in a slight performance drop

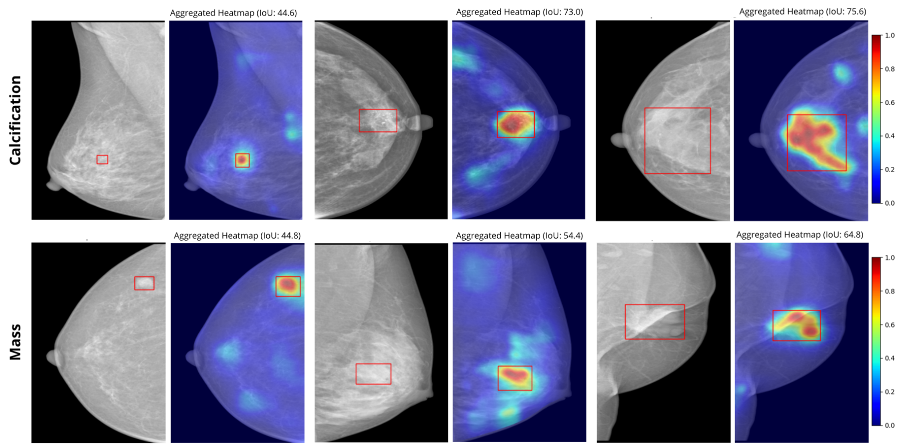

Fig. 2. Multi-scale aggregated heatmaps produced by the proposed FPN-MIL model, namely the FPN-SetTrans for calcifications and FPN-AbMIL for masses.

across most metrics for both lesion types, supporting prior findings on the benefits of multi-scale integration for model optimization [8, 17]. Conversely, feature concatenation of scale-specific bag embeddings (concat MS-Aggr) was the worst configuration regarding AUC and mAP metrics, particularly hindering lesion detection. While it achieves a comparable mAP_s and mAP_l but a significantly lower mAP_m for calcifications, for masses it actually achieves the highest $\text{mA}P_s$ and mAP_m but a drastically lower mAP_l. These results suggest ineffective feature fusion diluting discriminative information at specific scales, as already reported in the MIL literature [7, 14, 27]. Notably, the attention-based multi-scale aggregator achieves the best trade-off between classification and detection performances by adaptively weighting scale-specific features, more effectively preserving relevant information across scales and enhancing robustness to lesion scale variability.

Table 2. Ablation studies comparing different instance encoders (Inst-Enc) and multi-scale aggregators (MS-Aggr) for the best-performing FPN-MIL models. Detection performance is reported for all (mAP), small (mAP_s), medium (mAP_m) and large (mAP_l) lesions. The last row corresponds to our FPN-MIL models.

		Calcification					**Mass**				
Inst-Enc	**MS-Aggr**	**AUC**	**mAP**	**mAP$_s$**	**mAP$_m$**	**mAP$_l$**	**AUC**	**mAP**	**mAP$_s$**	**mAP$_m$**	**mAP$_l$**
MSP	Attention	91.3	18.5	0.3	22.8	54.9	77.1	9.5	0.0	9.5	46.6
FPN	w/o	93.8	33.0	8.5	35.7	61.6	78.8	25.2	5.0	30.7	56.0
FPN	Concat	92.2	28.8	12.6	17.2	59.4	76.9	19.4	**7.0**	**32.6**	26.4
FPN	Attention	**94.2**	**37.4**	**18.8**	**39.5**	**62.2**	**79.2**	**28.2**	4.7	32.1	**66.2**

5 Conclusion

In this work, we propose a novel weakly supervised FPN-MIL model for BC classification and detection, integrating an FPN-based instance encoder with multi-scale receptive-field granularity, deep-supervised scale-specific instance aggregators that support either AbMIL or SetTrans, and an attention-based multi-scale aggregator for a unified multi-scale analysis. Experimental results demonstrated that our FPN-MIL models significantly improves lesion detection over conventional single/multi-scale patch-based MIL models, with FPN-SetTrans performing best for calcifications and FPN-AbMIL for masses. In future work, we aim to extend our approach to end-to-end model training and explore other attention-based aggregators to further improve lesion detection under weak supervision.

Acknowledgments. This work was supported by FCT projects MIA-BREAST and AI-Radiologist (10.54499/2022.04485.PTDC and 10.54499/2024.07344.IACDC).

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

References

1. Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J.R., Cardoso, F., Siesling, S., Soerjomataram, I.: Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast **66**, 15–23 (2022)
2. Bobowicz, M., Rygusik, M., Buler, J., Buler, R., Ferlin, M., Kwasigroch, A., Szurowska, E., Grochowski, M.: Attention-based deep learning system for classification of breast lesions - multimodal, weakly supervised approach. Cancers **15**(10) (2023)
3. Buler, J., Buler, R., Bobowicz, M., Ferlin, M., Rygusik, M., Kwasigroch, A., Grochowski, M.: Interpretable deep learning approach for classification of breast cancer - a comparative analysis of multiple instance learning models. In: 2023 27th International Conference on Methods and Models in Automation and Robotics (MMAR). pp. 105–110 (2023)
4. Cheng, H., Shi, X., Min, R., Hu, L., Cai, X., Du, H.: Approaches for automated detection and classification of masses in mammograms. Pattern Recognition **39**(4), 646–668 (2006)
5. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis **54**, 280–296 (2019)
6. Choukroun, Y., Bakalo, R., Ben-Ari, R., Akselrod-Ballin, A., Barkan, E., Kisilev, P.: Mammogram Classification and Abnormality Detection from Nonlocal Labels using Deep Multiple Instance Neural Network. In: Eurographics Workshop on Visual Computing for Biology and Medicine. The Eurographics Association (2017)
7. Deng, R., Cui, C., Remedios, L.W., Bao, S., Womick, R.M., Chiron, S., Li, J., Roland, J.T., Lau, K.S., Liu, Q., Wilson, K.T., Wang, Y., Coburn, L.A., Landman, B.A., Huo, Y.: Cross-scale multi-instance learning for pathological image diagnosis. Medical Image Analysis **94**, 103124 (2024)

8. Early, J., Deweese, Y.J.C., Evers, C., Ramchurn, S.: Extending scene-to-patch models: Multi-resolution multiple instance learning for earth observation. Environmental Data Science **2**, e42 (2023)
9. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 632–642. Springer Nature Switzerland (2024)
10. Hu, T., Zhang, L., Xie, L., Yi, Z.: A multi-instance networks with multiple views for classification of mammograms. Neurocomputing **443**, 320–328 (2021)
11. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning (2018)
12. Lai, Y., Liu, X., E., L., Cheng, Y., Liu, S., Wu, Y., Zheng, W.: Transformer based multiple superpixel-instance learning for weakly supervised segmenting lesions of interstitial lung disease. Expert Systems with Applications **253**, 124270 (2024)
13. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 3744–3753. PMLR (2019)
14. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14313–14323 (2021)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017)
16. Lu, M.Y., Williamson, D.F.K., 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**, 555 – 570 (2020)
17. Marini, N., Otálora, S., Ciompi, F., Silvello, G., Marchesin, S., Vatrano, S., Buttafuoco, G., Atzori, M., Müller, H.: Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations. In: Proceedings of the MICCAI Workshop on Computational Pathology. Proceedings of Machine Learning Research, vol. 156, pp. 170–181. PMLR (2021)
18. Nguyen, H.T., Nguyen, H.Q., Pham, H.H., Lam, K., Le, L.T., Dao, M., Vu, V.: Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. medRxiv (2022), https://www.medrxiv.org/content/early/2022/03/10/2022.03.07.22272009
19. Oza, P., Sharma, P., Patel, S., Bruno, A.: A bottom-up review of image analysis methods for suspicious region detection in mammograms. Journal of Imaging **7**(9) (2021)
20. Qian, Z., Li, K., Lai, M., Chang, E.I.C., Wei, B., Fan, Y., Xu, Y.: Transformer based multiple instance learning for weakly supervised histopathology image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 160–170. Springer Nature Switzerland, Cham (2022)
21. Quintana, G.I., Li, Z., Vancamberg, L., Mougeot, M., Desolneux, A., Muller, S.: Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification. Bioengineering **10**(5) (2023)
22. Sarath, C.K., Chakravarty, A., Ghosh, N., Sarkar, T., Sethuraman, R., Sheet, D.: A two-stage multiple instance learning framework for the detection of breast cancer in mammograms. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1128–1131 (2020)

23. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., Geras, K.J.: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical Image Analysis **68**, 101908 (2021)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
25. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018)
26. Xie, L., Zhang, L., Hu, T., Huang, H., Yi, Z.: Neural networks model based on an automated multi-scale method for mammogram classification. Knowledge-Based Systems **208**, 106465 (2020)
27. Yoshida, T., Uehara, K., Sakanashi, H., Nosato, H., Murakawa, M.: Multi-scale feature aggregation based multiple instance learning for pathological image classification. In: International Conference on Pattern Recognition Applications and Methods. pp. 619–628 (2023)
28. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. bioRxiv (2016)
29. Zou, L., Yu, S., Meng, T., Zhang, Z., Liang, X., Xie, Y.: A technical review of convolutional neural network-based mammographic breast cancer diagnosis. Computational and Mathematical Methods in Medicine **2019**(1), 6509357 (2019)