

MIL-PF: Multiple Instance Learning on Precomputed Features for Mammography Classification

Nikola Jovišić^{1,2} (✉), Milica Škipina^{1,2,3}, Nicola Dall’Asen^{3,4,5}, and Dubravko Čulibrk^{2,6}

¹ Institute for AI R&D of Serbia, Novi Sad, Serbia
nikolajovisic4@gmail.com

² Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia

³ University of Trento, Trento, Italy

⁴ University of Pisa, Pisa, Italy

⁵ Fondazione Bruno Kessler, Trento, Italy

⁶ Jožef Stefan Institute, Ljubljana, Slovenia

Abstract. Modern foundation models provide highly expressive visual representations, yet adapting them to high-resolution medical imaging remains challenging due to limited annotations and weak supervision. Mammography, in particular, is characterized by large images, variable multi-view studies, and predominantly breast-level labels, making end-to-end fine-tuning computationally expensive and often impractical. We propose Multiple Instance Learning on Precomputed Features (MIL-PF), a scalable framework that combines frozen foundation encoders with a lightweight MIL head for mammography classification. By precomputing the semantic representations and training only a small task-specific aggregation module (40k parameters), the method enables efficient experimentation and adaptation without retraining large backbones. The architecture explicitly models the global tissue context and the sparse local lesion signals through attention-based aggregation. MIL-PF achieves state-of-the-art classification performance at a clinical scale while substantially reducing training complexity. We release the code for full reproducibility on <https://nikolajovisic.github.io/mil-pf/>.

Keywords: Breast Cancer · Multiple Instance Learning (MIL) · Frozen Encoders · Classification · Attention · Explainability

1 Introduction

Breast cancer is the most common malignancy and leading cause of cancer deaths in women [6]. The preferred screening and diagnostic method is mammography, and the analysis of mammograms requires considerable effort from radiologists, who have been increasingly relying on computer-aided diagnosis (CAD) [4].

Among clinical imaging modalities, mammography stands out in terms of high resolution (up to 4708×5844 pixels) [20] and typically lacks rich, pixel-level

annotations or scaled textual supervision. This limits CLIP-style training [18], often reducing it to synthetic captions, while standard CLIP-style setups hinder high-res and native multi-view modeling; as reflected in MammoCLIP [7]. Instead of training the encoders, we show that even when completely frozen, DINOv2 [14] and MedSigLIP [21] generalize well to mammography; this remains underexplored in prior work [5]. Freezing precludes encoder-focused approaches (LoRA [8], I-JEPA [1], etc.), but is more practical for weakly-supervised, multi-view, high-res inputs where encoder passes are costly. When the network head is trained in a fixed feature space, the embeddings can be precomputed, significantly reducing the total cost. As large research laboratories continue to focus on specialized encoders and sustainable AI, we anticipate that this approach will become even more valuable, particularly for research groups with limited resources.

To complete the pipeline, we frame the task using **Multiple Instance Learning (MIL)**. This allows us to adapt to realistic clinical scenarios where a single label is associated with a “bag” of instances (all views of a single breast). We explore this task in light of its hierarchical MIL structure, multiscale signal importance, and weakly-labeled regime. Unlike approaches that interleave attention layers with the encoder [2], our MIL head is trained independently. With only $\sim 40k$ trainable parameters, it is lightweight and easier to deploy in clinical settings. MIL-PF outperforms the closest prior approach [12], while using a simpler architecture and relying only on general encoders.

We evaluate MIL-PF on the mammography classification task (with visual explainability), but formalize a framework to make domain generalization accessible. Our main contributions are: 1. We formalize mammography classification as a specific MIL problem and propose a novel arrangement of components designed for it. 2. We leverage recent advances in foundation vision models to explore feature precomputing and frozen encoders as a principled design choice for mammography MIL, showing that strong general backbones enable expressive pipelines without the need for end-to-end fine-tuning. 3. We validate the approach at clinical scale, demonstrate state-of-the-art performance on mammography classification benchmarks, and provide code to enable full reproducibility.

2 Method

Multiple Instance Learning (MIL). In Multiple Instance Learning (MIL) [3,25,9], the dataset $\mathcal{D}$ is composed of T labeled bags, as defined in Eq. (1).

$$\mathcal{D} = \{(\mathcal{S}_i, y_i)\}_{i=1}^T, \quad \text{where } \mathcal{S}_i = \{I_i^{(n)}\}_{n=1}^{N_i} \text{ and } y_i \in \mathcal{L} \quad (1)$$

Each bag $\mathcal{S}_i$ is a collection of N_i instances $I_i^{(n)}$. The label $y_i \in \mathcal{L}$ is a bag-level label that applies to the entire bag $\mathcal{S}_i$; the labels for individual instances $I_i^{(n)}$ are unknown. The learning task is to map each bag $\mathcal{S}_i$ to its corresponding label y_i .

Problem Formulation. We study a special case of MIL, which includes:

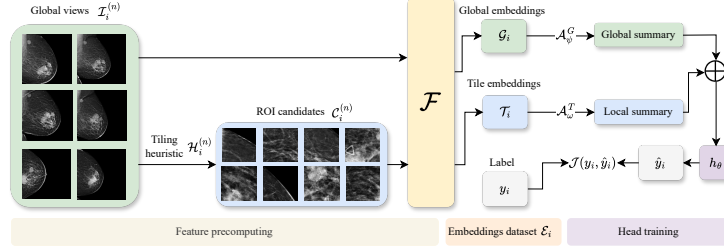


Fig. 1. The MIL-PF pipeline diagram, divided into distinct stages: Feature Precomputing producing Embeddings dataset $\mathcal{E}$ and Head Training using them.

1. **Nested hierarchy:** Each instance $I_i^{(n)}$ from a bag $\mathcal{S}_i$ is a high-resolution image. This image, in turn, contains a variable number of sparsely distributed Regions of Interest (ROIs), $\{R_i^{(n)(k)}\}$. Crucially, these ROIs are not annotated.
2. **Complementary streams:** The images $\{I_i^{(n)}\}_n$ within a bag convey a comparable low-frequency signal (global stream), e.g. pertinent to general tissue structure, while substantial, fine-grained differences exist only at high resolution within their respective ROIs (local stream).

Embeddings Dataset. MIL methods have traditionally been shown to have problems with backpropagation, especially with sparsely present task-relevant instances [9]. Multiscale reasoning, as required by our problem, hinders it further.

Therefore, we first transform our data using a pretrained encoder. We assume the existence of an encoder, $\mathcal{F}$, that is pretrained to expressively encode both full images $I_i^{(n)}$ and high-resolution sub-regions, the tiles, in a zero-shot setting. Using this encoder, we generate two distinct sets of embeddings for each bag $\mathcal{S}_i$.

First, we create a set of global embeddings, $\mathcal{G}_i$, by applying the encoder to each full image in the bag, as in Eq. (2).

$$\mathcal{G}_i = \{\mathcal{F}(I_i^{(n)})\}_n \quad (2)$$

Second, to capture the local signals, we must extract features from the unannotated ROIs. We create candidates for these ROIs, $\{C_i^{(n)(k)}\}_{k=1}^{M_i^{(n)}}$, by dividing each image $I_i^{(n)}$ into a uniform grid of tiles. The heuristic $\mathcal{H}_i^{(n)}$ we use for selecting the tiles is taking all containing any breast tissue from the grid of encoder-input-sized tiles, discarding only the fully background ones. The tile size must be large enough to fit the expected ROI, but not larger than the maximum resolution the encoder supports. The number of tiles $M_i^{(n)}$ may vary per image.

We then create a single set of tile embeddings, $\mathcal{T}_i$, by encoding all candidate tiles from all images in the bag and collecting them together, as in Eq. (3).

$$\mathcal{T}_i = \bigcup_n \bigcup_{k=1}^{M_i^{(n)}} \{\mathcal{F}(C_i^{(n)(k)})\} \quad (3)$$

Finally, as defined in Eq. (4), we form a unified embeddings dataset, $\mathcal{E}$, which consists of the global embeddings, the tile embeddings, and the corresponding bag label y_i for each bag in our original dataset.

$$\mathcal{E} = \{(\mathcal{G}_i, \mathcal{T}_i, y_i)\}_i \quad (4)$$

MIL-PF Head. For simplicity, our model employs a late-fusion strategy of the two streams, as on Fig. 1. Even though learning complex interactions of the global $\mathcal{G}_i$ and the local $\mathcal{T}_i$ stream could in principle be useful, it is not crucial for our task. We define inference in Eq. (5).

$$\hat{y}_i = h_\theta(\text{concat}(\mathcal{A}_\psi^G(\mathcal{G}_i), \mathcal{A}_\omega^T(\mathcal{T}_i))), \quad (\mathcal{G}_i, \mathcal{T}_i, y_i) \in \mathcal{E} \quad (5)$$

- $\mathcal{A}_\psi^G$ and $\mathcal{A}_\omega^T$ are permutation-invariant aggregators with parameters ψ and ω that produce a single summary vector for the global and local embedding sets, respectively. Prior to actual pooling operations, they pass the embeddings through a small Multi-Layer Perceptron (MLP). As the encoder is frozen, this part does the actual processing of individual crops with respect to the task, and this is where the majority of the trainable parameters are.
- h_θ is a final layer that maps the joint summary to a prediction $\hat{y}_i$.

The training objective minimizes the loss $\mathcal{J}(y_i, \hat{y}_i)$ with respect to (θ, ψ, ω) .

A key advantage of this design is that the pretrained encoder $\mathcal{F}$ performs most of the heavy lifting. The MIL-PF module itself can therefore be lightweight with a low parameter count, focusing solely on correctly implementing the domain’s inductive biases. As shown in Section 3, this modularity provides numerous benefits: fast experiment iteration, drastically lower computational requirements and the ease of swapping encoders or inserting additional inductive biases.

Attention Pooling for Local Embeddings. A careful choice of the local aggregator ($\mathcal{A}_\omega^T$) is necessary, as the task-relevant ROIs are sparse in the set $\mathcal{T}_i$.

Standard pooling methods are suboptimal for this task. Mean pooling fails, as the signal from a few important tiles is suppressed or “diluted” by the overwhelming number of unimportant background tiles, and max pooling can only capture information from a single, most-salient tile, even if multiple ROIs are relevant for the bag-level prediction. An attention mechanism is therefore preferred. However, common candidates like self-attention are ill-suited, as they are designed to primarily model the *interdependence* between inputs, while we are interested just in attending to relevant tiles.

Therefore, we employ a **Perceiver-style cross-attention** mechanism [10] as our local aggregator $\mathcal{A}_\omega^T$. This approach uses a single, trainable latent vector z that acts as a query. This latent query z attends to the set of all local tile embeddings from $\mathcal{T}_i$, which are projected into Keys (K) and Values (V). The mechanism computes a weighted sum of the Values based on the query-key similarity ($\text{softmax}(zK^T)V$), which “pulls” the most relevant information from the tile set into a single summary vector. This design is highly parameter-efficient,

focusing the model’s capacity on learning the weights of the aggregators and the final head. In our experiments, we found that using a single latent query was sufficient and that adding more latents (as in the standard Perceiver) did not provide additional benefits for this task.

3 Experiments and Results

3.1 Experimental Setup

Datasets. We evaluate on one of the largest public mammography datasets, the EMBED Open Data, which contains data for ~ 0.5 M mammograms [11], representing highly diverse real-world clinical scenarios [24]. For comparison with other mammography approaches, we also evaluate on VinDr [13] and RSNA [19].

A bag $\mathcal{S}_i$ is defined as a set of all views of a single breast in a given exam. We use breast-wise labels, as opposed to image-wise, because radiologists form reports considering all of the views together. Patient-wise reasoning, on the other hand, collapses ("max-pools") to the label of the more suspicious breast, losing a valuable training signal. We define R_i as regions containing lesions (annotating these is very hard to scale due to the cost and highly diverse pathologies).

Even the weak case-level label assignment strategy (BI-RADS reports) is noisy and varies among radiologists. Therefore, for general malignancy risk modeling evaluation, we use BI-RADS 1 as negative; 4, 5 and 6 as positive, across mixed clinical cohorts; and discard 2 and 3. The split varies widely in the literature and this one minimizes noise and ensures fair statistical comparison, while retaining principal clinical significance. Using solely BI-RADS 6 ground truth is statistically unreliable, since those examples are scarce even in the largest datasets.

Metrics. The metrics used in the literature vary. We use **AUC** for class separability; **bACC** to reflect performance regarding class imbalance; **mAP** for lesion detection; and Spec@Sens=0.9 (abb. **Spec.**), specificity at a threshold that yields a sensitivity of 0.9. We chose the last metric due to clinical relevance of sensitivity [17]; one is often concerned about how much specificity can be achieved with a high sensitivity constraint. We note that it does not strictly correlate with AUC.

Implementation Details. We split the data 70%-10%-20% for train, validation and test sets, balanced for all used BI-RADS values proportionally, assuring no patient leakage across splits. The MLP part of the $\mathcal{A}^G$ and $\mathcal{A}^T$ consist of two layers followed by ReLU activations, the first layer projecting from embedding size (encoder-specific) to 16, and second from 16 to 8; we experimentally observed these hyperparameters to be optimal. The loss function used is Binary Cross-Entropy. For every experiment, all bags of each split of the whole embeddings dataset $\mathcal{E}$, even for the large EMBED dataset, can fit in a single batch of our A100 40GB GPU and we train in such a setting, requiring ~ 5 -7 min for a single run of full MIL-PF head training and ~ 2 M FLOPS for a forward pass per breast. For classification, we do not use overlapping tiles, as we observed no

Table 1. Encoder selection (left) and ablation on aggregation (right) on EMBED. † indicates aggregation applied at inference only.

Model	Resolution	AUC ↑	Spec. ↑	Aggregation		DINOv2	ViT Giant	MedSigLIP	
				$\mathcal{A}^G$	$\mathcal{A}^L$			AUC ↑	Spec. ↑
MammoCLIP [7]	1520 × 912	0.870	0.558	mean†	—	0.871	0.627	0.870	0.619
BiomedCLIP [26]	224 × 224	0.872	0.606	mean	—	0.878	0.640	0.873	0.640
MedSigLIP [21]	448 × 448	0.897	0.691	max†	—	0.865	0.601	0.873	0.621
RADDINO [16]	518 × 518	0.854	0.499	max	—	0.897	0.665	0.897	0.691
DINOv2 [14] ViT Giant	518 × 518	0.897	<u>0.655</u>	attn	—	0.896	0.668	0.899	0.676
DINOv3 [23] ViT Huge+	512 × 512	0.831	0.497	—	attn	0.892	0.662	0.899	0.713
DINOv3 [23] ViT Huge+	1024 × 1024	0.834	0.476	max	mean	0.901	0.705	0.902	0.672
				max	max	0.905	0.703	0.918	0.735
				max	attn	<u>0.916</u>	0.762	0.914	<u>0.746</u>

improvement in performance. For the attention map computation we use 75% overlap at inference time. IoU threshold for mAP is 0.25.

Multiple Runs. Since the head training is inexpensive with the precomputed embeddings, we run 36 independent runs per each head training experiment. Although we observe some variance across runs (up to 2% AUC and 11% Spec.), selecting the model with the highest validation AUC consistently yields near-maximal test performance. Thus, the results reflect a reliable optimization effect rather than a single fortunate run.

3.2 Encoder Selection

Training a simplified version of MIL-PF on EMBED BI-RADS task that relies just on max-pooling-based $\mathcal{A}^G$ gives us a general overview of the backbone choices. As shown in Table 1 (left), DINOv2 [14] and MedSigLIP [21] overperform other medical and general encoders, as well as the mammography-pretrained MammoCLIP [7], which does not generalize as well to a new dataset.

3.3 Comparison with the Baselines

We compare MIL-PF against recent state-of-the-art models that address a similar task and have publicly available code for reproduction: Shen *et al.* [22], Pathak *et al.* [15] and Mourão *et al.* [12]. As the competing models are not breast-level-aligned, we additionally evaluate the main competitor (Mourão *et al.*) using mean and max aggregation over multiple views. All the models were retrained on the same data splits as MIL-PF on VinDr, RSNA and EMBED for fair comparison.

Table 2 and Table 3 show MIL-PF’s competitive or superior performance, using only a fraction of competitors’ trainable parameters (total expressed in millions as Par.), especially on the largest scale and most diverse EMBED dataset. We do not attribute this advantage to any single component of our method by itself, but to a combination of factors: MIL-PF being the first in the literature to discover generalization properties of the general encoders to the mammography domain;

this further enabling feature precomputing, fast iterative experimentation and multiple runs optimization; and finally resulting in a domain-aligned aggregation mechanism (ablated in Section 3.5) with inherent breast-wise modeling.

Table 2. Quantitative comparison of MIL-PF against SOTA baselines on VinDr and EMBED on BI-RADS-based malignancy. MIL-PF demonstrates superior performance, particularly on the large-scale, noisy EMBED dataset, highlighting its robustness.

Model/Paper	Par. Level	VinDr			EMBED		
		AUC $\uparrow$	bAcc $\uparrow$	Spec. $\uparrow$	AUC $\uparrow$	bAcc $\uparrow$	Spec. $\uparrow$
ES-Att ^{side} [15]	22.89 Case	0.786	0.749	0.193	0.836	0.668	0.558
SIL ^{IL} GMIC-ResNet18 [15]	22.49 Image	<u>0.911</u>	0.855	0.698	0.875	0.802	0.566
GMIC [22]	14.11 Image	0.899	0.826	0.690	0.816	0.736	0.380
FPN-AbMIL [12]	1.76 Image	0.920	0.844	<u>0.720</u>	0.802	0.740	0.367
FPN-SetTrans [12]	5.38 Image	0.908	0.827	0.670	0.794	0.728	0.364
FPN-AbMIL (mean) [12]	1.76 Breast	<u>0.911</u>	0.831	0.643	0.835	0.765	0.403
FPN-AbMIL (max) [12]	5.38 Breast	0.899	0.830	0.512	0.817	0.759	0.334
FPN-SetTrans (mean) [12]	1.76 Breast	0.900	0.811	0.559	0.835	0.767	0.451
FPN-SetTrans (max) [12]	5.38 Breast	0.887	0.776	0.595	0.813	0.728	0.388
MIL-PF (DINOv2) $\mathcal{A}^T$: max	0.05 Breast	0.882	0.811	0.712	0.905	0.833	0.703
MIL-PF (DINOv2) $\mathcal{A}^T$: attn	0.05 Breast	0.894	0.845	0.792	<u>0.916</u>	0.850	0.762
MIL-PF (MedSigLIP) $\mathcal{A}^T$: max	0.04 Breast	0.901	0.838	0.616	0.918	<u>0.845</u>	0.735
MIL-PF (MedSigLIP) $\mathcal{A}^T$: attn	0.04 Breast	<u>0.911</u>	<u>0.850</u>	0.683	0.914	0.843	<u>0.746</u>

Table 3. Quantitative comparison of MIL-PF against SOTA baselines on the VinDr and RSNA datasets. MIL-PF demonstrates superior performance.

Model/Paper	Par. Level	VinDr - Mass			VinDr - Calcification			RSNA - Cancer		
		AUC $\uparrow$	bAcc $\uparrow$	Spec. $\uparrow$	AUC $\uparrow$	bAcc $\uparrow$	Spec. $\uparrow$	AUC $\uparrow$	bAcc $\uparrow$	Spec. $\uparrow$
ES-Att ^{side} [15]	22.89 Case	0.738	0.669	0.266	0.937	0.864	0.877	0.890	0.799	0.642
SIL ^{IL} GMIC-ResNet18 [15]	22.49 Image	0.772	0.703	0.307	0.953	0.886	0.886	0.899	0.754	0.684
GMIC [22]	14.11 Image	0.755	0.687	0.286	0.941	0.855	0.899	0.773	0.699	0.434
FPN-AbMIL [12]	1.76 Image	0.792	0.720	0.382	0.954	0.902	0.904	0.891	0.793	0.628
FPN-SetTrans [12]	5.38 Image	0.785	0.734	0.367	0.942	0.884	0.836	0.874	0.773	0.656
FPN-AbMIL (mean) [12]	1.76 Breast	<u>0.808</u>	0.757	0.387	0.962	0.913	0.916	0.914	0.826	0.687
FPN-AbMIL (max) [12]	5.38 Breast	0.797	0.683	0.332	<u>0.958</u>	0.899	0.911	0.902	0.785	0.664
FPN-SetTrans (mean) [12]	1.76 Breast	0.807	<u>0.755</u>	0.317	0.952	0.881	0.867	0.902	0.825	<u>0.747</u>
FPN-SetTrans (max) [12]	5.38 Breast	0.801	0.745	0.342	0.947	0.877	0.892	0.896	0.773	0.693
MIL-PF (DINOv2) $\mathcal{A}^T$: max	0.05 Breast	0.770	0.731	0.311	<u>0.958</u>	<u>0.917</u>	0.884	0.901	0.815	0.669
MIL-PF (DINOv2) $\mathcal{A}^T$: attn	0.05 Breast	0.800	0.736	0.419	0.967	0.930	0.931	<u>0.923</u>	0.828	0.709
MIL-PF (MedSigLIP) $\mathcal{A}^T$: max	0.04 Breast	0.778	0.720	<u>0.408</u>	0.967	0.907	0.861	0.925	<u>0.830</u>	0.733
MIL-PF (MedSigLIP) $\mathcal{A}^T$: attn	0.04 Breast	0.814	0.757	0.403	0.967	0.835	<u>0.917</u>	<u>0.923</u>	0.835	0.757

3.4 Explainability Analysis

Table 4 reports detection performance across different lesion sizes (as defined in [12]) on VinDr. MIL-PF achieves competitive results for medium and large sizes, while mAP for small lesions remains low. However, qualitative results on

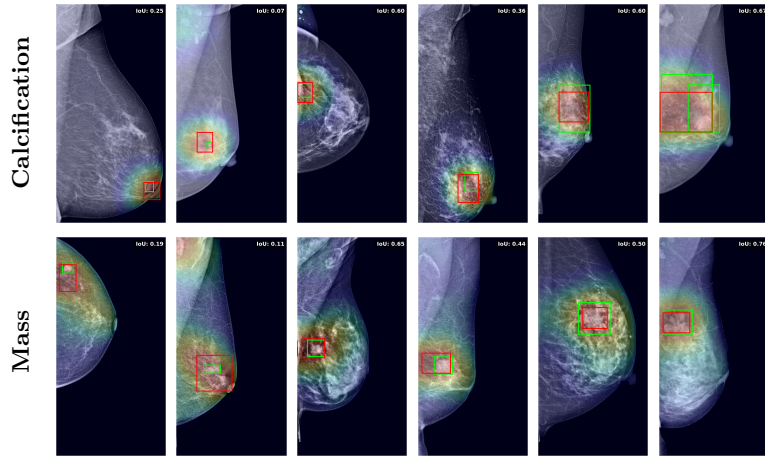


Fig. 2. Attention maps for masses and calcifications obtained by overlapping the MIL-PF’s local stream window. Ground truth is shown in green and predictions in red.

randomly selected examples in Fig. 2 show that our models mostly struggle with reaching IoU@0.25 threshold, while consistently identifying the correct principal regions. We attribute this to the relatively large input tile sizes (448×448 / 518×518), which, while not hindering leading classification performance, affect detection resolution. We consider this a possibility for future enhancement.

Table 4. Comparison of proposed MIL-PF with FPN[12]. Detection performance is reported for all (mAP), small (mAP_s), medium (mAP_m) and large (mAP_l) lesions.

Model/Paper	Calcification				Mass			
	mAP ↑	mAP _s ↑	mAP _m ↑	mAP _l ↑	mAP ↑	mAP _s ↑	mAP _m ↑	mAP _l ↑
FPN-AbMIL [12]	32.0	9.1	34.8	57.5	28.2	4.7	32.1	66.2
FPN-SetTrans [12]	37.4	18.8	39.5	62.2	24.3	3.0	28.0	73.2
MIL-PF (DINOv2)	24.4	0.4	22.5	94.44	16.3	0.1	15.9	62.8
MIL-PF (MedSigLIP)	38.4	1.2	52.5	84.9	14.1	0.1	11.3	58.4

3.5 Ablation Studies

To quantify the significance of each component of MIL-PF, we train differently configured MIL architectures. The results are shown in Table 1 (right).

The configurations containing $\mathcal{A}^G$ use global features, the ones containing $\mathcal{A}^T$ use tile features, and the ones containing both, use both, as in full MIL-PF. The aggregations labeled with † are only performed at the inference level, while for the training each image inherits its bag’s label. This simulates non-MIL approaches, but aggregation is still preserved at inference to preserve consistency.

We note that the difference between the full MIL-PF setup and naïve single instance learning (SIL) performed only on global views goes up to 5% for AUC and 14% for Spec., proving MIL-PF’s inductive biases correct and its capacity capable of leveraging powerful backbones.

4 Conclusion and Future Work

We introduced MIL-PF, a lightweight design for weakly-labeled mammography data that operates on features precomputed by frozen foundation models. Our key finding is that a powerful general-purpose encoder, combined with a small specialized $\sim 40k$ parameter task-specific head, yields state-of-the-art results on a large-scale clinical dataset. Results challenge the prevailing assumption that end-to-end fine-tuning or complex custom architectures are prerequisites for this task. This opens multiple venues for future research, including the integration of more complex inductive biases, such as prior patient exams or bilateral asymmetry impact, and the application to other high-resolution, weakly-labeled domains.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15619–15629 (2023)
2. Chen, H., Martel, A.L.: Breast cancer detection from multi-view screening mammograms with visual prompt tuning. In: Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care. pp. 91–100. Springer Nature Switzerland, Cham (2026)
3. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* **89**(1-2), 31–71 (1997)
4. Doi, K.: Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics* (2007)
5. Germani, E., Selin-Türk, I., Zeineddine, F., Mourad, C., Albarqouni, S.: Bias and Generalizability of Foundation Models across Datasets in Breast Mammography. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. LNCS, vol. 15973, pp. 24–34. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-032-05185-1_3
6. Ghoncheh, M., Pournamdar, Z., Salehiniya, H.: Incidence and mortality and epidemiology of breast cancer in the world. *Asian Pacific Journal of Cancer Prevention* (2016)
7. 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. LNCS, vol. 15012, pp. 632–642. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_59

8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)* (2022)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *International Conference on Machine Learning (ICML)*. PMLR (2018)
10. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: *International Conference on Machine Learning (ICML)* (2021)
11. Jeong, J.J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilie, G., Smith, G., et al.: The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiology: Artificial Intelligence* **5**(1), e220047 (2023)
12. Mourão, M., Nascimento, J.C., Santiago, C., Silveira, M.: Multi-scale attention-based multiple instance learning for breast cancer diagnosis. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. LNCS, vol. 15974, pp. 364–374. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-032-05182-0_36
13. 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. *Scientific Data* (2023)
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)* (2024)
15. Pathak, S., Schlötterer, J., Geerdink, J., Veltman, J., van Keulen, M., Strisciuglio, N., Seifert, C.: Breast cancer prediction using mammography exams for real hospital settings. *Computers in Biology and Medicine* (2025)
16. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence* **7**, 119–130 (2025). <https://doi.org/10.1038/s42256-024-00965-w>
17. Petitti, D.B., Penhoet, E.E., Joy, J.E.: Saving women’s lives: strategies for improving breast cancer detection and diagnosis. National Academies Press (2005)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning (ICML)*. pp. 8748–8763. PMLR (2021)
19. Radiological Society of North America (RSNA): RSNA screening mammography breast cancer detection (RSNA-SMBC) dataset (2024), accessed: January 26, 2024
20. Rangarajan, K., Gupta, A., Dasgupta, S., Marri, U., Gupta, A.K., Hari, S., Banerjee, S., Arora, C.: Ultra-high resolution, multi-scale, context-aware approach for detection of small cancers on mammography. *Scientific Reports* (2022)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: MedGemma technical report. *arXiv preprint arXiv:2507.05201* (2025)
22. Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., et al.: An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis* (2021)

23. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
24. Woo, M., Zhang, L., Brown-Mulry, B., Hwang, I., Gichoya, J.W., Gastounioti, A., Banerjee, I., Seyyed-Kalantari, L., Trivedi, H.: Subgroup evaluation to understand performance gaps in deep learning-based classification of regions of interest on mammography. PLOS Digital Health (2025)
25. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
26. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. NEJM AI **2**(1) (2025). <https://doi.org/10.1056/AIoa2400640>