Supplementary Material

A. Confusion Matrix for FPN-IAIA-BL

Figure 6 contains the confusion matrix for FPN-IAIA-BL on the test set. It has the highest specificity and lowest sensitivity for the circumscribed class.

TARGET / OUTPUT	Circumscribed	Indistinct	Spiculated	SUM
Circumscribed	14 17.95%	1 1.28%	1 1.28%	16 87.50% 12.50%
Indistinct	8 10.26%	30 38.46%	3 3.85%	41 73.17% 26.83%
Spiculated	3 3.85%	3 3.85%	15 19.23%	21 71.43% 28.57%
SUM	25 56.00% 44.00%	34 88.24% 11.76%	19 78.95% 21.05%	59 / 78 75.64% 24.36%

Figure 6. Confusion matrix for predictions on the test dataset.

B. Fine Annotation Coefficients

FPN-IAIA-BL introduces fine-annotation coefficients $\lambda_{\text{in}}^{(y(i),c)}, \lambda_{\text{out}}^{(y(i),c)}$ which are used in the fine-annotation loss to encourage and penalize the model for activating inside and outside the fine annotations. For example, it is considered "worse" for a spiculated prototype to activate on a circumscribed lesion than for a circumscribed prototype to activate on a spiculated lesion. The fine-annotation coefficients designed by board-certified radiologist F.S. are as follow in tables 2 and 3.

		Prototype Class			
		Circ.	Ind.	Spic.	Neg.
Sample's Class	Circ.	1	1	1	1
	Ind.	1	1	1	1
	Spic.	1	1	1	1
	Neg.	0	0	0	0

Table 2. Fine annotation coefficients penalizing the prototypes from class c_{proto} from activating **outside** fine annotations for a sample from class c_i

Incorporating the fine-annotation coefficients, the fine-annotation loss is now defined as:

$$\ell_{\text{fine}} = \sum_{i \in D'} \sum_{\mathbf{p}^{(c,l,j)}} \Big(||\lambda_{\text{in}}^{(y(i),c)} \mathbf{m}_i \odot \text{PAM}_{i,j} + \lambda_{\text{full}}^{(y(i),c)} \text{PAM}_{i,j}||_2 \Big) \tag{5}$$

		Prototype Class			
		Circ.	Ind.	Spic.	Neg.
Sample's Class	Circ.	0	0	0	1
	Ind.	0	0	0	1
	Spic.	1	1	0	1
	Neg.	0	0	0	0

Table 3. Fine annotation coefficients penalizing the prototypes from class c_{proto} from activating **inside** the fine annotations for a sample from class c_i.

where the prototype activation map $\text{PAM}_{i,j}$ is computed by bilinearly upsampling the similarity map $[s_{j,n}]_{n=(1,1)}^{(\eta_l,\eta_l)}$ for prototype $\mathbf{p}^{(c,l,j)}$ and image $\mathbf{x}_i$ such that it has the same dimensions as the fine-annotation mask $\mathbf{m}_i$.

C. Negative Class Data

As discussed in Section 4, we include a negative class during training to discourage "classification by elimination." The negative class data consist of 5,000 image-mask pairs. The negative class images were created by sampling the full-size mammogram images and cropping to a section without any of the lesion region of interest for each image. We pair each image with a fully negative mask where no region of interest is identified in the mask. For training, we randomly select a subset of 200 negative samples.