Calibration in Multiple Instance Learning: Evaluating Aggregation Methods for Ultrasound-Based Diagnosis

Axel Geysels[1 (✉)] [0000−0002−5922−1030], Ben Van Calster[1] [0000−0003−1613−7450], Bart De Moor[1] [0000−0002−1154−5028], Wouter Froyman[1,2] [0000−0002−1398−9124], and Dirk Timmerman[1,2] [0000−0002−3707−6645]

[1] KU Leuven, Leuven, Belgium
axel.geysels@esat.kuleuven.be
[2] UZ Leuven, Leuven, Belgium

Abstract. Accurate probability estimates are critical for clinical decision-making, yet many Multiple Instance Learning (MIL) methods prioritize classification performance alone. We investigate the calibration quality of various MIL aggregation strategies, comparing them against simpler instance-based probability pooling in both *in-distribution* and *out-of-distribution* ultrasound imaging scenarios. Our findings reveal that attention-based aggregators yield stronger discrimination but frequently produce overconfident predictions, leading to higher calibration errors. In contrast, simpler instance-level methods offer more reliable risk estimates, albeit with a modest reduction in classification metrics. These results underscore a trade-off between predictive strength and calibration in MIL, emphasizing the importance of evaluating both aspects for clinically robust applications.

Keywords: Multiple Instance Learning · Calibration · Ultrasound

1 Introduction

Multiple Instance Learning (MIL) is especially well-suited to medical imaging scenarios in which each patient (bag) is represented by multiple images (instances) yet only receives a single, high-level label (e.g., benign vs. malignant) [15]. This setup naturally arises in clinical contexts such as histopathology, mammography, and ultrasound, where per-image annotations are costly and time-consuming, or simply infeasible [6, 2]. MIL frameworks address this challenge by aggregating instance-level features into a single bag-level prediction and can additionally highlight the most influential images—an important property for interpretability and clinical decision-making.

Despite extensive research on improving the *discriminative* performance of MIL systems, relatively little attention has been paid to the *calibration* of their predicted probabilities. Calibration refers to how well a model's predicted risks align with true event frequencies—a critical aspect of reliability in medical settings. For instance, overconfident estimates can lead to unnecessary interventions, whereas underestimation may delay essential care [13]. Ensuring properly calibrated predictions is thus paramount, yet the calibration characteristics of different MIL aggregation methods remain largely unexplored.

In this work, we investigate the calibration of several MIL aggregation strategies—ranging from simple mean pooling to more sophisticated attention-based mechanisms, and we benchmark them against instance-level baselines in both in-distribution and out-of-distribution ultrasound imaging scenarios.

2 Methodology

2.1 Multiple Instance Learning (MIL) Setup

We consider a multiple instance learning (MIL) framework where each patient (bag) b has a collection of 2D transvaginal ultrasound (TVUS) images (instances):

$$X^b = \{\mathbf{x}_1^b, \mathbf{x}_2^b, \ldots, \mathbf{x}_{N_b}^b\}, \tag{1}$$

with N_b the number of images in bag b. We assume the images in each bag are independent, have no inherent ordering, and that N_b can vary per patient. There are B patients in total, with $b \in \{1, \ldots, B\}$. A single binary label $Y^b \in \{0, 1\}$ is assigned to each bag, indicating benign ($Y^b = 0$) or malignant ($Y^b = 1$). Under the classical MIL assumption, each instance within the bag also has an (unobserved) binary label $y_i^b \in \{0, 1\}$. The bag-level label relates to the instance-level labels by:

$$Y^b = \max\{y_1^b, \ldots, y_{N_b}^b\}, \tag{2}$$

hence $Y^b = 1$ if and only if *at least one* instance is positive.

2.2 MIL Aggregation

Each image $\mathbf{x}_i^b \in \mathbb{R}^{3HW}$ (with 3 denoting color channels, and H, W the spatial dimensions) is passed through a backbone to produce a feature vector $\mathbf{h}_i^b \in \mathbb{R}^d$. We then *pool* or *aggregate* the set of instance embeddings $\{\mathbf{h}_1^b, \ldots, \mathbf{h}_{N_b}^b\}$ into a single bag-level embedding $\mathbf{H}^b$. We explore four MIL pooling techniques:

Mean Pooling — We compute the simple average over instance features:

$$\mathbf{H}^b \;=\; \frac{1}{N_b} \sum_{k=1}^{N_b} \mathbf{h}_k^b. \tag{3}$$

Max Pooling — We take the per-dimension maximum across instances:

$$(\mathbf{H}^b)_m \;=\; \max_{1 \leq k \leq N_b} (\mathbf{h}_k^b)_m, \quad m = 1, \ldots, d. \tag{4}$$

Attention-Based Aggregation — Following Ilse et al. [9], we adopt a gated attention mechanism, where each instance embedding $\mathbf{h}_k^b \in \mathbb{R}^d$ is mapped into an L-dimensional space:

$$\mathbf{A}_k^V = \tanh\bigl(V\,\mathbf{h}_k^b\bigr), \quad \mathbf{A}_k^U = \sigma\bigl(U\,\mathbf{h}_k^b\bigr), \tag{5}$$

where $U, V \in \mathbb{R}^{L\times d}$. We combine them elementwise:

$$\mathbf{A}_k^{\text{comb}} = \mathbf{A}_k^V \odot \mathbf{A}_k^U, \quad a_k = \mathbf{w}^\top \mathbf{A}_k^{\text{comb}}, \tag{6}$$

followed by a softmax to produce attention weights α_k. The final bag-level embedding is the weighted sum of instance embeddings:

$$\mathbf{H}^b = \sum_{k=1}^{N_b} \alpha_k\,\mathbf{h}_k^b.$$

Attention- and Uncertainty-Based Aggregation —

We propose an inverse-entropy re-weighting of the gated-attention scores that softly down-weights high-entropy (i.e. uncertain) instances while retaining the full set of images in the bag. A conceptually related idea—"Certainty Pooling"—was introduced by Gildenblat et al. [8], performing a hard arg-max selection that relies on Monte-Carlo dropout, whereas our variant keeps the original soft attention and requires only a single deterministic forward pass. After the standard attention weights α_k are computed, the binary-class entropy of instance k is

$$E_k \;=\; -\Bigl[p_k\,\log p_k \;+\; (1-p_k)\,\log(1-p_k)\Bigr]. \tag{7}$$

with p_k the confidence that instance k belongs to the bag-level label Y^b. We define an inverse-uncertainty weight

$$u_k \;=\; \frac{1}{1+E_k}, \tag{8}$$

multiply each α_k by u_k, and re-normalize to again ensure the weights sum to 1:

$$\tilde{\alpha}_k \;=\; \frac{\alpha_k \cdot u_k}{\sum_{j=1}^{N_b} \alpha_j\,u_j}. \tag{9}$$

Finally, the bag embedding becomes

$$\mathbf{H}^b \;=\; \sum_{k=1}^{N_b} \tilde{\alpha}_k\,\mathbf{h}_k^b. \tag{10}$$

2.3 Calibration of Bag-Level Predictions

Let $\widehat{Y}$ be the model's predicted label and $\widehat{P}$ the associated confidence, or the probability of correctness. We aim for *calibrated* predictions, meaning that when the model outputs a confidence p, the true frequency of positives should be p. Formally, calibration requires:

$$\mathbb{P}(\widehat{Y} = Y \mid \widehat{P} = p) \;=\; p, \quad \forall p \in [0,1]. \tag{11}$$

While perfect calibration is seldom attained in practice, it is commonly assessed via *calibration plots (also known as reliability diagrams* and the *expected calibration error (ECE)*. The *Brier score* is also frequently reported; influenced by calibration, it is fundamentally an overall performance metric because it also depends on discrimination [1].

Calibration Plots — A common way to visualize calibration is to group predictions into M bins (e.g. deciles) and plot the average confidence (predicted probability) against the observed accuracy (event rate) in each bin [4, 12]. Let B_m be the set of indices of patients whose prediction confidence falls into the interval $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right)$. The observed accuracy within B_m is then:

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{b \in B_m} \mathbf{1}(\widehat{y}^b = y^b), \tag{12}$$

where $\widehat{y}^b$ and y^b are the predicted and true class labels for patient (bag) b. The average confidence for bin B_m is:

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{b \in B_m} (\widehat{p}^b), \tag{13}$$

where $\widehat{p}^b$ is the confidence of patient b. This bin-based approach helps reveal whether predicted risks align with actual frequencies.

An alternative is to fit a *logistic calibration model*, as often proposed in clinical risk prediction modeling [14, 3]:

$$\mathit{logit}\big(\mathbb{P}(\widehat{Y}^b = 1 \mid \widehat{P}^b = p^b)\big) \;=\; \alpha \;+\; \zeta\, \mathit{logit}\big(p^b\big). \tag{14}$$

Perfect calibration requires $\alpha = 0$ and $\zeta = 1$. A slope $\zeta < 1$ means predicted probabilities are too extreme, whereas $\zeta > 1$ suggests they are too moderate. When focusing solely on the intercept, we fix $\zeta = 1$ so that the resulting calibration intercept α' reveals whether the predicted risks are overestimated ($\alpha' < 0$) or underestimated ($\alpha' > 0$) on average.

Expected Calibration Error (ECE) — Calibration Plots are primarily visual, whereas Expected Calibration Error (ECE) [11] offers a scalar metric of miscalibration, more precisely it is defined as the difference in expectation between confidence and accuracy, given by:

$$\mathrm{ECE} \;=\; \sum_{m=1}^{M} \frac{|B_m|}{B} \Big| \mathrm{acc}(B_m) \;-\; \mathrm{conf}(B_m) \Big|, \tag{15}$$

where $\{B_m\}_{m=1}^{M}$ are bins of predictions, B is the total number of patients, $\mathrm{acc}(B_m)$ is the observed proportion of positives, and $\mathrm{conf}(B_m)$ the average predicted probability in bin m. Lower ECE values indicate better calibration.

Brier Score — Finally, the Brier Score (BS) [7] measures the mean squared difference between predicted probabilities and true class labels:

$$\mathrm{BS} \;=\; \frac{1}{B} \sum_{b=1}^{B} (\widehat{p}^{b} - y^{b})^{2}, \tag{16}$$

As a proper scoring rule, BS rewards well-calibrated predictions (the lower, the better).

3 Experiments

Dataset We used an interim *Prospective Ultrasound Ovarian Tumor* dataset consisting of 8,824 images from 1,457 patients, collected between 2019 and 2024 across 11 centers in 6 countries. Each patient's histological diagnosis (benign vs. malignant) served as the binary label. Overall, 37.2% of bags (patients) were malignant, whereas 40.2% of instances (images) were malignant. The difference arises because some malignant patients had more images, thereby increasing the proportion of malignant images. The number of images per patient ranged from 1 to 30 (Mean = 6.06, SD = 5.26). Figure 1 presents the distribution of patients per center, bundling centers contributing fewer than 5% of patients.

Implementation Details We employed the *ConvNeXt-small* architecture [10] (ImageNet-22k pre-train, ImageNet-1k fine-tune [5]) as the backbone for all models, followed by a single linear classifier. Images are resized to 224 × 224, z-score normalized, and augmented with standard affine and elastic transforms. Models are trained with the stochastic gradient descent (SGD) optimizer with momentum, a learning rate of 5×10^{-4}, and a cosine annealing scheduler. For the *MIL variants*, we formed mini-batches that pack several complete bags until the total reaches at most 64 images—well above the maximum of 30 images per patient. Cross-entropy is computed per bag, then averaged across the mini-batch, yielding a single scalar for back-propagation while ensuring efficient GPU utilization. For the *instance baselines*, the backbone produces a probability for each image individually; patient-level scores are then obtained by mean or max pooling over that set of image-level probabilities.

Evaluation We evaluated performance using two strategies: an *in-distribution (ID)* approach with 5-fold cross-validation at the patient level, and an *out-of-distribution (OOD)* approach based on leave-one-center-out cross-validation following the center scheme in Figure 1. For each fold or held-out center, we measured validation set performance, after which we pooled the validation set predictions across folds (or centers) into a single set, to compute the overall performance metrics with 95% confidence intervals derived via bootstrapping. Classification metrics included accuracy, F1-score (both thresholded at 0.5), and the area under the ROC curve (AUC). Calibration was examined with the expected calibration error (ECE) and logistic calibration plots. Additionally, we report the Brier score, providing an indicator of overall model performance.

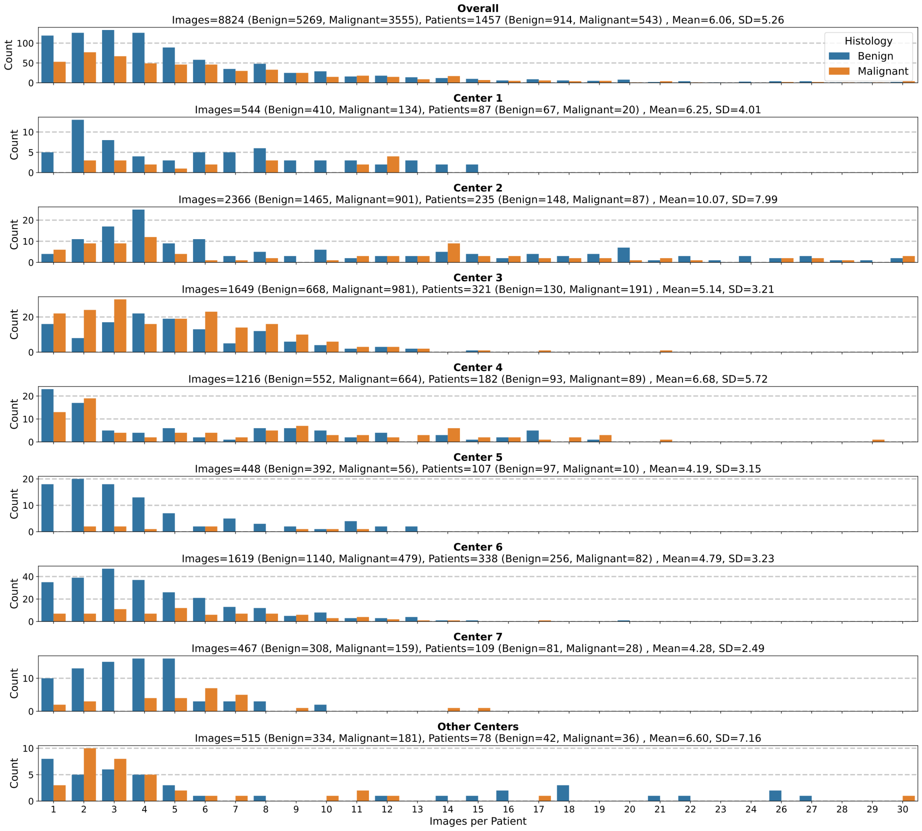

Fig. 1. Distribution of images per patient at each center, stratified by histology. Centers contributing fewer than 5% of all patients are grouped under *Other Centers*.

4 Results and Discussion

Table 1 summarizes classification and calibration metrics for two validation protocols: in-distribution (ID), using 5-fold cross-validation at the patient level, and out-of-distribution (OOD), via leave-one-center-out cross-validation. All metrics use the *raw* output probabilities—no post-hoc calibration method was used—such that any differences arise solely from the aggregation strategy itself. Under ID conditions, MIL+GA+Uncertainty shows a modest improvement in AUC and F1-score compared to Instance+Mean; however, Instance+Mean offers the best calibration (lowest ECE and Brier). Notably, MIL+Max also achieves relatively low ECE. In the OOD scenario, MIL+GA obtains the highest accuracy and F1-score, while MIL+GA+Uncertainty yields the highest AUC. Even so, Instance+Mean again demonstrates superior calibration.

Figure 2 provides the logistic calibration plots (both per fold/center and aggregated) alongside prediction histograms. These highlight that all MIL-based

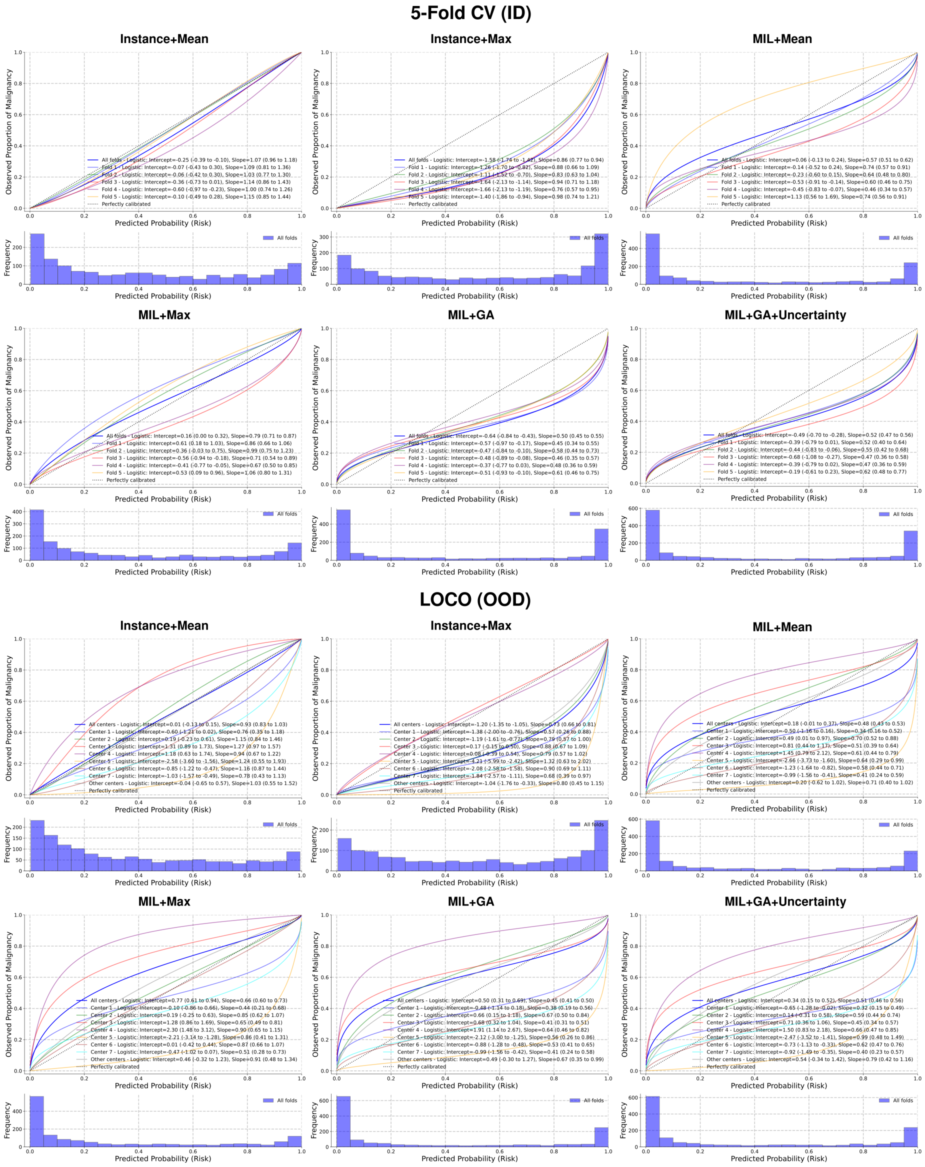

Fig. 2. Logistic calibration plots (both per fold/center and aggregated) and corresponding prediction histograms for different MIL aggregation strategies, evaluated under in-distribution (ID) and out-of-distribution (OOD) validation protocols.

Table 1. Performance metrics (with 95% confidence intervals) aggregated for both the patient-wise 5-fold (ID) and leave-one-center-out (OOD) validation strategies.

Validation	Method	ACC	F1	AUC	ECE	Brier
5-Fold CV (ID)	Instance+Mean	0.828 (0.808-0.846)	0.768 (0.741-0.795)	0.905 (0.889-0.920)	**0.033** (0.026-0.054)	**0.120** (0.110-0.130)
	Instance+Max	0.774 (0.753-0.795)	0.753 (0.729-0.779)	0.904 (0.888-0.920)	0.172 (0.155-0.191)	0.164 (0.151-0.177)
	MIL+Mean	0.823 (0.802-0.842)	0.759 (0.729-0.788)	0.893 (0.877-0.910)	0.069 (0.057-0.091)	0.131 (0.118-0.144)
	MIL+Max	**0.830** (0.811-0.849)	0.759 (0.730-0.788)	0.895 (0.877-0.911)	0.036 (0.027-0.058)	0.124 (0.112-0.136)
	MIL+GA	0.820 (0.800-0.840)	0.769 (0.740-0.796)	0.906 (0.890-0.921)	0.087 (0.075-0.109)	0.132 (0.119-0.146)
	MIL+GA+Uncertainty	0.826 (0.805-0.846)	**0.773** (0.745-0.799)	**0.908** (0.892-0.922)	0.093 (0.077-0.112)	0.131 (0.118-0.145)
LOCO (OOD)	Instance+Mean	0.800 (0.778-0.820)	0.717 (0.684-0.748)	0.863 (0.844-0.883)	**0.020** (0.018-0.045)	**0.142** (0.131-0.153)
	Instance+Max	0.747 (0.725-0.770)	0.714 (0.685-0.742)	0.861 (0.842-0.882)	0.144 (0.127-0.166)	0.175 (0.162-0.188)
	MIL+Mean	0.797 (0.777-0.818)	0.719 (0.688-0.750)	0.867 (0.849-0.888)	0.100 (0.082-0.119)	0.151 (0.136-0.166)
	MIL+Max	0.802 (0.781-0.822)	0.693 (0.658-0.726)	0.871 (0.853-0.891)	0.086 (0.070-0.106)	0.146 (0.132-0.159)
	MIL+GA	**0.806** (0.786-0.828)	**0.724** (0.691-0.756)	0.871 (0.852-0.890)	0.106 (0.091-0.128)	0.153 (0.137-0.168)
	MIL+GA+Uncertainty	0.799 (0.778-0.821)	0.718 (0.685-0.749)	**0.872** (0.854-0.892)	0.103 (0.087-0.125)	0.150 (0.134-0.165)

strategies tend to push bag-level probabilities closer to 0 or 1, thus producing overconfident predictions—albeit less so for MIL+Max. Overall, while attention-based approaches can offer modest gains in classification performance, they typically come at the cost of poorer calibration. Meanwhile, simpler instance-level methods (e.g., Instance+Mean) maintain more reliable probability estimates but yield slightly lower discrimination.

Because each patient has only a single histology label, we lack image-level ground truth and cannot compute quantitative instance-level metrics (e.g., an instance-level AUC). We therefore examine attention patterns qualitatively. Figure 3 presents a malignant case with five images: the instance baseline outputs a separate probability for every image, whereas MIL+GA+Uncertainty yields one bag-level probability plus attention weights. The first, fourth, and fifth images receive the highest weights—aligning with visible solid tissue suspicious for malignancy—demonstrating how attention can surface clinically meaningful cues and offer an interpretable rationale behind model predictions.

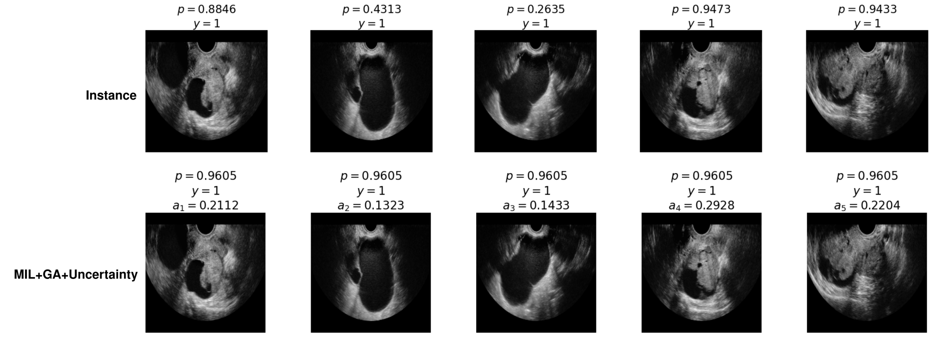

Fig. 3. Comparison of instance-level (per-image) predictions versus a gated attention-based MIL approach that outputs a single bag-level probability along with per-image attention weights. Shown here is a malignant bag, where the first, fourth, and fifth images receive higher attention weights, corresponding to visibly solid tissue regions, indicative of malignancy.

Future Steps and Key Takeaways. Our results reveal a trade-off between discrimination and calibration in MIL. While attention-based approaches may boost classification performance, they often produce overconfident predictions. Moving forward, calibration metrics should be a central consideration in developing new MIL frameworks, particularly in medical imaging domains where poorly calibrated risk estimates may have serious clinical consequences.

Acknowledgments. This work is supported by Research Foundation Flanders (FWO), KU Leuven Internal Funds, and the Flanders AI Research Program. Data collection: Giulia Garofalo, Stefan Timmerman, Francesca Buonomo, Daniela Fischerová, Antonia Carla Testa, Francesca Moro, Lil Valentin, Povilas Sladkevicius, Ligita Jokubkiene, Marek Kudla, Caroline Van Holsbeke, Artur Czekierdowski, Elisabeth Epstein, and Yvette Groszmann.

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

References

1. Assel, M., Sjoberg, D.D., Vickers, A.J.: The brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagnostic and Prognostic Research **1**(1), 19 (2017)
2. Barbosa, D., Ferreira, M., Junior, G.B., Salgado, M., Cunha, A.: Multiple instance learning in medical images: A systematic review. IEEE Access **12**, 78409–78422 (2024)
3. Campo, B.D.C.: Towards reliable predictive analytics: a generalized calibration framework (2023)
4. DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. Journal of the Royal Statistical Society. Series D (The Statistician) **32**(1), 12–22 (1983)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
6. Gadermayr, M., Tschuchnig, M.: Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. Computerized Medical Imaging and Graphics **112**, 102337 (2024)
7. Gerds, T.A., Cai, T., Schumacher, M.: The performance of risk prediction models. Biometrical Journal **50**(4), 457–479 (2008)
8. Gildenblat, J., Ben-Shaul, I., Lapp, Z., Klaiman, E.: Certainty pooling for multiple instance learning. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I. p. 141–153. Springer-Verlag, Berlin, Heidelberg (2021)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (2018)
10. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s (2022)

11. Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. p. 2901–2907. AAAI'15, AAAI Press (2015)
12. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. p. 625–632. ICML '05, Association for Computing Machinery, New York, NY, USA (2005)
13. Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative: Calibration: the achilles heel of predictive analytics. BMC Med. **17**(1), 230 (2019)
14. Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M.J., Steyerberg, E.W.: A calibration hierarchy for risk models was defined: from utopia to empirical data. J. Clin. Epidemiol. **74**, 167–176 (2016)
15. Waqas, M., Ahmed, S.U., Tahir, M.A., Wu, J., Qureshi, R.: Exploring Multiple Instance Learning (MIL): A brief survey. Expert Systems with Applications **250**, 123893 (2024)