Robust sensitivity control in digital pathology via tile score distribution matching

Arthur Pignet[1,†], John Klein[1], Geneviève Robin[1,*], and Antoine Olivier[1,*,†]

[1] Owkin, Inc
† Corresponding author.
* Equal senior authorship.
{arthur.pignet, john.klein, genevieve.robin, antoine.olivier}@owkin.com

Abstract. Deploying digital pathology models across medical centers is challenging due to distribution shifts. Recent advances in domain generalization improve model transferability in terms of aggregated performance measured by the Area Under Curve (AUC). However, clinical regulations often require to control the transferability of other metrics, such as prescribed sensitivity levels. We introduce a novel approach to control the sensitivity of whole slide image (WSI) classification models, based on optimal transport and Multiple Instance Learning (MIL). Validated across multiple cohorts and tasks, our method enables robust sensitivity control with only a handful of calibration samples, providing a practical solution for reliable deployment of computational pathology systems.

Keywords: Digital pathology · Multiple instance learning · Sensitivity Control · Optimal transport

1 Introduction

Deep learning (DL) models are used in Computational Pathology (CPath) to analyze whole slide images (WSI) in a variety of medical contexts [15,14,19]. However, their deployment in the clinic is limited by their ability to generalize well beyond the training context, impaired by inherent data variability due to the use of different scanners, staining and labeling protocols [20]. To overcome this, recent works in DL for CPath have applied domain generalization (DG) techniques, designed to increase the robustness of predictive models to distribution shifts between training and evaluation [21,2,10,12,25,1,8].

The performance of DG methods is usually evaluated with the Area Under the ROC Curve (AUC), which may fail to capture their generalization capacity in clinical contexts, where sensitivity and specificity are also crucial [13]. Indeed, the sensitivity/specificity trade-off is often controlled via a threshold transforming continuous scores into binary labels. Even when models preserve the AUC, the distribution of predicted scores may vary between cohorts [18], impairing the threshold's capacity to yield consistent sensitivity and specificity [7], as illustrated in Figure 1. To solve this issue, existing works have relied on calibration procedures to adjust score distributions.

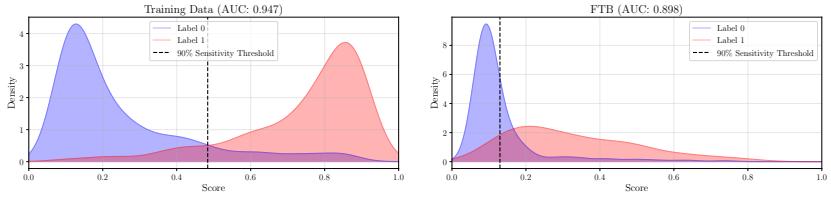


Fig. 1: Distribution of a DL model's prediction scores on the training (top) and validation cohorts (FTB, bottom) with comparable AUC (train: 94.7%, val: 89.8%). Dashed line correspond to the threshold associated to a sensitivity of 90.0%, (train: 0.45, FTB: 015).

Strictly speaking, probabilistic calibration refers to the transformation of prediction scores into actual class membership probabilities [6], for instance using temperature scaling [11]. While probabilistic model calibration is a way to control sensitivity, it is a stronger requirement aiming to control sensitivity at *any level*. In practice, a weaker requirement is that models achieve *a fixed sensitivity level* prescribed by the clinical context (say, 90%). Thus, aligned with existing work, we use 'calibration' in its more general meaning, referring to a model's capacity to output similar scores at training and inference time. In practice, this is often achieved using calibration data from the deployment center, to adjust the model's threshold after training [19,18].

In this paper, we develop a new methodology called Tile-Score Matching (TSM) to control the sensitivity in WSI binary classification problems. We focus on Multiple Instance Learning (MIL) models based on the Chowder architecture [5], which has been applied to many WSI classification problems [19,23,24,4]. In Chowder, each WSI is divided into tissue patches called tiles; prediction scores are computed at the tile-level, then aggregated into a single WSI-level score. We leverage this property by working at the tile level to calibrate the distribution of prediction scores. TSM is most similar to Unsupervised Prediction Alignment (UPA) [18], which also calibrates the distribution of prediction scores. However, while UPA matches score distributions at the WSI level, TSM operates at the tile level, increasing the number of available calibration samples by several orders of magnitude. As a result, while UPA requires hundreds of WSI for calibration, TSM requires less than 30.

Our contributions are presented as follows. In Section 2, we introduce TSM, a new threshold calibration method which matches the distribution of tile-level prediction scores to a reference distribution using optimal transport (OT), and accounts for prevalence shift with importance sampling (IS). We also provide theoretical evidence of sensitivity control in particular cases. In Section 3, we demonstrate empirically that TSM controls the sensitivity across several indications and classification tasks[1]. Our experiments also show that, contrary to

[1] Our code, model and features are available at https://github.com/owkin/tsm.git.

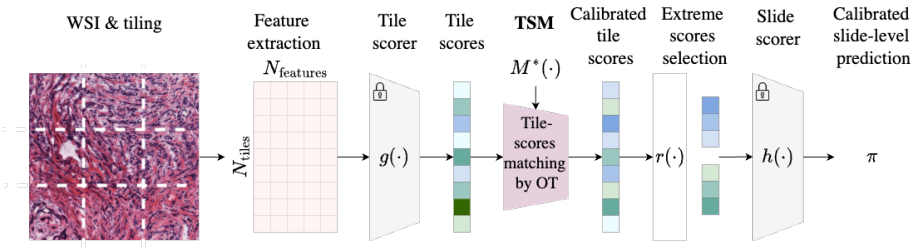


Fig. 2: Graphical overview of the method.

existing works, TSM controls sensitivity in extremely low data and prevalence regimes, when only 5 positive samples are available for calibration.

2 Tile-Score Distribution Matching

In this section, we present our method and explain how it can be combined with Chowder [5] to target pre-specified sensitivity levels, as illustrated in Figure 2. For clarity, we restrict to sensitivity, but the extension of the method to reach pre-specified specificity levels is straightforward by simply considering negative samples instead of positive samples.

General framework. Let $\mathbb{S}$ denote the observation space of WSI, S^t a WSI from the reference cohort (possibly used during training) and Y^t the associated binary label. Respectively, denote S^c a WSI from the calibration population and Y^c the associated binary label. In what follows, we use the superscript $\{t, c\}$ to denote the membership to the training or calibration cohorts. Let $(\mathbb{S}\times\{0,1\}, \mathcal{S}\otimes\mathcal{P}\{0,1\}, \mu)$ be a measurable space, and assume that (S^t, Y^t) and (S^c, Y^c) admit positive density functions with respect to the base measure μ.

Similarly to other MIL models, Chowder identifies a WSI $S \in \mathbb{S}$ to a collection of tiles belonging to S. For simplicity, we assume that each WSI has the same number of tiles N, and denote $S = (T_1, \ldots, T_N)$, $T_i \in \mathbb{T} := \mathbb{R}^{d\times d}$ for all $1 \leq i \leq N$ (where $d \times d$ is the tile size). By design, the prediction function f is the composition of a tile-level scoring function $g : \mathbb{T} \mapsto \mathbb{R}$, a selection function $r : \mathbb{R}^N \mapsto \mathbb{R}^{2k}$, which ranks and selects the top-k and bottom-k tile-scores, and a predictor $h : \mathbb{R}^{2k} \mapsto \mathbb{R}$:

$$f(S) = h(r(g(T_1), \ldots, g(T_N))). \tag{1}$$

Finally, the predicted patient label is given by $\mathbb{1}\{f(S) \geq \tau\}$. Usually, τ is adjusted post-training to achieve the prescribed sensitivity level σ,

$$\mathbb{P}_{S,Y}\left(f(S) > \tau | Y = 1\right) = \sigma, \tag{2}$$

by leveraging held-out calibration data (see, *e.g.*, the methodology deployed in [19]). As previously highlighted, τ often does not transfer to external cohorts,

in the sense that the achieved level of sensitivity will drift apart from the prescribed target σ due to distributional shifts. For instance, in Figure 1, $\tau = 0.45$ is calibrated during training to achieve $\sigma_{\text{train}} = 0.9$, but yields $\sigma_{\text{val}} = 0.32$ on the validation cohort. We develop the TSM methodology to mitigate this lack of transferability, by matching the distribution of tile scores $g(T_i)$ between the training and application domains.

Tile-score distribution matching. By construction, for $1 \leq i \leq N$ and for $k \in \{t, c\}$, T_i^k admits a positive density function on $\mathbb{T}$. For simplicity of exposition, we assume that, in each cohort, tiles are i.i.d. conditionally to the slide label. Thus, for $1 \leq i, j \leq N$, $k \in \{t, c\}$ and for $l \in \{0, 1\}$, $\mathbb{P}(T_i^k | Y^k = l) = \mathbb{P}(T_j^k | Y^k = l)$. This assumption is required to prove the theoretical control of sensitivity, yet, it may not hold in practice, as adjacent tissue samples have correlated spatial structure. We note however that the sparsity of the tiling can be controlled, for instance by extracting only a subset of tiles, serving as an effective way to reduce the spatial correlation between the considered tiles. Besides, we show experimentally in section 3 that sensitivity is controlled in real-life scenarios.

As the tile score function g is continuous, the random variables $(X_i := g(T_i^k))_{1 \leq i \leq N}$ are also i.i.d. conditionally to the label Y^k. Denoting $\omega^k = \mathbb{P}(Y^k = 1)$ the prevalence in cohort k, the density function of a tile score X is given for $k \in \{c, t\}$ by

$$\rho_X^k = \omega^k \rho_{X|Y=1}^k + (1 - \omega^k) \rho_{X|Y=0}^k. \tag{3}$$

We now explain how the tile score distribution ρ_X^c is matched, up to an adjustment w.r.t. prevalence, to the reference distribution ρ_X^t. The Monge formulation of OT writes, for two measures a and b on $\mathbb{R}$:

$$M^* = \underset{M \in \mathcal{M}_{a \to b}}{\arg\min} \int_{\mathbb{R}} |M(x) - x|^2 \mathrm{d}a(x), \tag{4}$$

where $\mathcal{M}_{a \to b}$ is the set of Borel measurable functions such that $M_{\#}a = b$, *i.e.*, b is the push-forward measure of a through M. Since X is one-dimensional, (4) has a closed-form solution, which is the monotonous map obtained through quantile matching [16]. The optimal map is given by $M^* = F_b \circ {F_a}^{-1}$, where F_a and F_b are the cumulative distribution functions of measures a and b and ${F_a}^{-1}$ is the generalized inverse of F_a. TSM consists in applying (4) to $a = \rho_X^c$ and

$$b = \omega^c \rho_{X|Y=1}^t + (1 - \omega^c) \rho_{X|Y=0}^t, \tag{5}$$

a reweighted version of ρ_X^t which accounts for prevalence shift between training and validation. We now prove that TSM controls the sensitivity level when $\omega^c = 1$, *i.e.*, when the calibration set contains only positive labels.

Theorem 1. *Let $\tau \in \mathbb{R}$, and denote by $sens_{train}(\tau)$ and $sens_{val}(\tau)$ the sensitivities associated to threshold τ on the training and validation cohorts. Assume that the calibration set contains only positive examples, i.e., $\omega^c = 1$. Then,*

$$sens_{val}(\tau) = sens_{train}(\tau). \tag{6}$$

Proof. After calibration, the sensitivity on the validation cohort is given by $\text{sens}_{val}(\tau) = \mathbb{P}_{M^*\#(\rho_X^c)}(f(S^c) > \tau | Y = 1)$. By construction of the transport map M^*, and with $\omega^c = 1$, we have $M^*_\#(\rho_X^c) = \rho^t_{X|Y=1}$ and $\rho^c_{X|Y=1} = \rho_X^c$. Thus,

$$\begin{aligned}
\text{sens}_{val}(\tau) &= \mathbb{P}_{M^*_\#(\rho_X^c)}(f(S^c) > \tau | Y = 1) \\
&= \int_{f^{-1}(]\tau,1])} M^*_\#(\rho_X^c)(x_1) \dots M^*_\#(\rho_X^c)(x_N) dx_1 \dots dx_N \\
&= \int_{f^{-1}(]\tau,1])} \rho^t_{X|Y=1}(x_1) \dots \rho^t_{X|Y=1}(x_N) dx_1 \dots dx_N \\
&= \mathbb{P}(S^t \in f^{-1}(]\tau, 1])|Y = 1) = \mathbb{P}(f(S^t) > \tau | Y = 1) = \text{sens}_{train}(\tau).
\end{aligned}$$

Theorem 1 implies that TSM, applied to a calibration set drawn conditionally on $Y = 1$, effectively ensures transferability of the model's sensitivity. However, as shown in section 3, using a calibration set containing both positive and negative samples can lead to similar sensitivity transfers.

Lemma 1. *Let $M^*_d : R^d \mapsto R^d$ be the d multi-dimensional (component-wise) application of M^* and r the ranking function of the Chowder model. Then,*

$$M^*_N \circ r = r \circ M^*_{2k}. \tag{7}$$

Proof. The function r in Chowder is a ranking function [4]. Thus, the monotonicity of M^* concludes the proof.

Lemma 1 states that the set of $2k$ tiles selected by the ranking layer of the Chowder model is invariant through calibration by TSM. This means in particular that interpretation properties are preserved, as the regions of the WSI used by the final prediction function h are unchanged. Lemma 1 also has computational implications, as the Monge map only needs to be applied after the ranking layer r, reducing the number of operations from N to $2k$.

In practice, the densities ρ_X^c and ρ_X^t are unknown, and we approximate them by the weighted sum of Dirac masses centered on each data point, i.e.,

$$\widehat{\rho_X^c} = \frac{1}{n_c} \sum_{i=1}^{n_c} \delta_{x^c_{(i)}}, \quad \text{and} \quad \widehat{\rho_X^t} = \frac{1}{n_t} \sum_{j=1}^{n_t} \delta_{x^t_{(j)}}, \tag{8}$$

where $(x^c_{(i)})_{1 \le i \le n_c}$ and $(x^t_{(j)})_{1 \le j \le n_t}$ are the empirical tile scores obtained by applying g to the training and validation set of tiles, respectively. Consequently, $F_{\widehat{\rho_X^c}}$ and $F_{\widehat{\rho_X^t}}$ are staircase functions. To provide more flexibility in M^*, we resort to a linear interpolation between function jumps.

3 Experiments

ER, PR and HER2 status prediction in breast cancer. In a first study, we predict the status of estrogen receptor (ER), progesterone receptor (PR) and

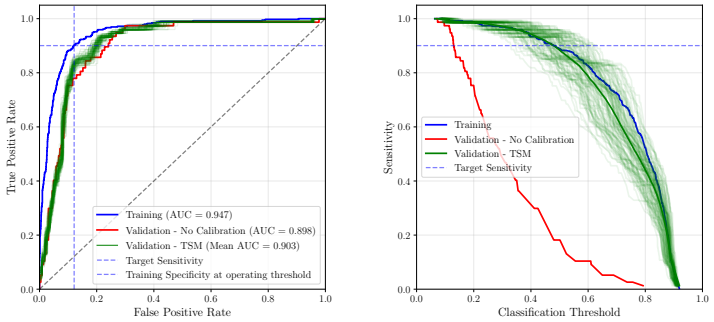
human epidermal growth factor receptor 2 status (HER2) in breast cancer. We use TCGA-BRCA as training cohort. For ER and PR, we use the 1076 labeled slides. The prevalence of ER+ (resp. PR+) patients in the training dataset is 78% (resp. 68%). For the HER2 task, we use the reliable labels from [17], yielding 801 slides (15% HER2+). Two external datasets are used for validation. For the 3 endpoints, we use the BCNB dataset [22] with early breast cancer core-needle biopsies collected from 1,058 patients, along with ER, PR and HER2 statuses (79% ER+, 75% PR+, 26% HER2+). Additionally, for HER2 status prediction, we use the Herohe dataset [3] (41% HER2+).

MSI status prediction in colorectal cancer. In a second study, we predict high Microstatellite instability (MSI-H) against low instability or stability (MSI-L or MSS). The training set is the combination of TCGA-COAD, TCGA-READ, and a private dataset from Medipath laboratories (France). The prevalence of MSI-H in the training dataset is 18%. Five external datasets are used for validation. Cypath is a private collection of 698 H&E and H&E&S biopsies from 698 patients (36% MSI-H) digitized in France; Cypath-HE and Cypath-HES cohorts are obtained by splitting Cypath to account for the variations in staining conditions. Neogenomics is a second private collection of 198 biopsies and 200 resections (43% overall); it also further splits into Neogenomics-resections and Neogenomics-biopsies. FTB is a private collection of 602 patients (26% MSI-H).

Training setup. For the 4 prediction tasks, a pre-trained feature extractor is first used to extract low-dimensional representations for tiles. We use Phikon, for the 3 breast-related tasks and a feature extractor tailored for colorectal cancer for MSI status prediction, as in [19]. Chowder models are then trained to predict a slide-level score with repeated cross-validation (3 repeats and 5 splits). For each split, 5 Chowder models are trained, resulting in an ensemble of 75 models. During validation, TSM is applied to each model independently, and the resulting 75 calibrated models are ensembled to produce the final prediction.

TSM controls sensitivity and preserves ROC curves. Figure 3a depicts the ROC curves with (in green) and without (in red) TSM calibration, and shows that TSM does not impact the ability of the model to rank patients correctly; Table 1 also highlights that the AUC is preserved by TSM. Figure 3b illustrates the good transfer of the sensitivity/threshold curve from the training cohort (in blue) to the calibrated external validation cohort (in green). On the contrary, the sensitivity/threshold curve without calibration (in red) shows that applying the training threshold to the external cohort without calibration would lead to a dramatic drop of sensitivity (from 90% to 20%).

Comparison with prior art. We compare TSM to two sensitivity control methods from the literature: UPA [18], discussed in Section 1, and Patient level threshold selection (PLTS, see [19]). PLTS leverages a calibration set composed



(a) ROC curves. (b) Sensitivity vs thresholds.

Fig. 3: Metrics computed on both training and FTB validation cohorts for the MSI classification task, with and without TSM. While the validation ROC curves maintain their distinct profiles, the calibrated sensitivity curves converge towards the training curve, suggesting effective calibration. Each shaded green curve represents an individual sampling of the calibration set.

Table 1: Transfer of model AUC across different cohorts. The training AUC is estimated via cross-validation; the validation AUC is computed on external cohorts, with and without TSM calibration.

Task	ER	PR	HER2	HER2	MSI Status				
Ext. Cohort	BCNB			HEROHE	Cypath		Neogenomics		FTB
					HE	HES	Bio.	Res.	
Train	0.904	0.809	0.818	0.818	0.947				
Validation	0.818	0.816	0.643	0.569	0.917	0.899	0.947	0.977	0.898
Calibrated Val.	0.825	0.818	0.633	0.604	0.913	0.898	0.958	0.977	0.903

of m WSIs with *constant labels*, denoted $(S_i^c)_{i=1}^m$, and matches the threshold τ to a quantile of the scores computed on the calibration data. In the case of sensitivity control, and if $\pi_i^c = f(S_i^c)$ denote the ordered calibration scores, i.e., $\pi_1^c \leq .. \leq \pi_m^c$, , it sets $\tau = \pi_i^c$ where the integer i is set to the largest such that $sens\,(\pi_i^c) \geq \sigma$. This methodology can be used with positive samples to control sensitivity (PLTS+) or with negative samples to control specificity (PLTS−). In Figure 4, we report the sensitivities obtained by all methodologies using 30 WSI for calibration. We observe that TSM and PLTS+ reach the desired sensitivity while keeping relatively close to the target, while UPA falls short in this low data regime. TSM and PLTS+ have similar average sensitivity, but TSM exhibits less variability.

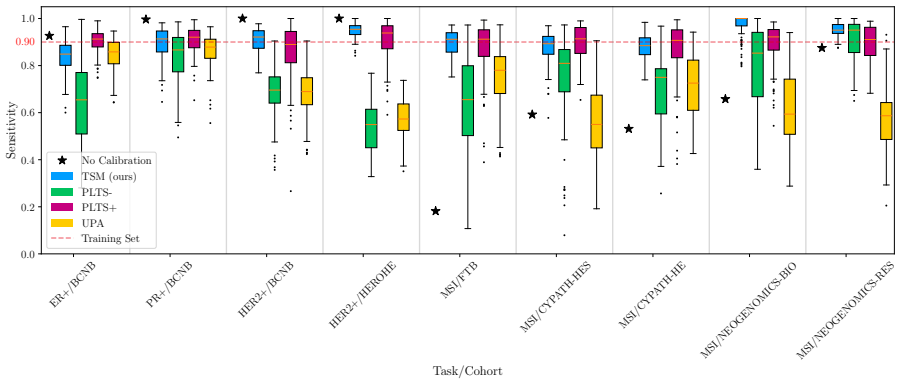


Fig. 4: Sensitivity achieved by TSM, UPA, and PLTS+/- on several tasks and validation cohorts using 30 slides for calibration, across 100 experiments.

TSM outperforms PLTS in low prevalence regime. In Figure 5, we further compare TSM and PLTS+ in the more challenging setting where only a handful of positive samples can be used for calibration. In this experiment, we focus on the low prevalence tasks with less than 30% positive samples, and PLTS+ and TSM are allowed only 5 positive samples for calibration. By design, PLTS+ can only use positive samples; on the contrary, TSM is able to also leverage the negative samples. Enriched with negative samples, TSM (5/20) improves over PLTS+, both in terms of targeted sensitivity and variability.

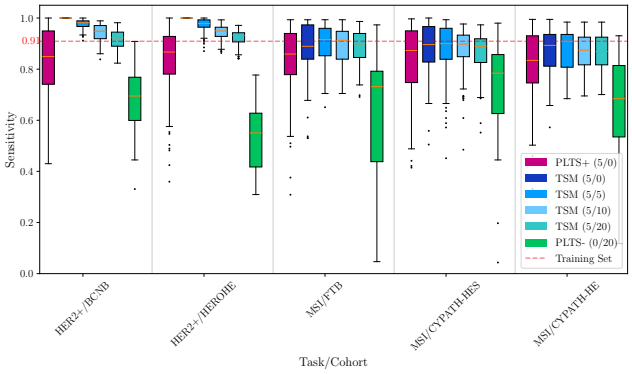


Fig. 5: Sensitivity achieved by PLTS+ and TSM using 5 positive samples for calibration. For TSM, we add up to 20 additional negative samples; the ratio (positive/negative) is provided in the legend.

4 Conclusion

In this paper, we propose a novel threshold calibration method to control sensitivity in WSI binary classification tasks. Using only a handful of calibration samples, TSM outperforms existing works in terms of sensitivity control on several indications and classification tasks. The main shortcoming of TSM is its specificity to the Chowder architecture which, although frequently used for WSI analysis, remains limiting. Thus, future work includes the extension of TSM to other MIL modeling approaches [9]. Another direction of future research is the analysis of TSM in settings where the exact prevalence of positive samples is unknown. For instance, in the context of the deployment of TSM in a new hospital, the historical prevalence of the center can be used, as well as the indication prevalence from existing epidemiological data or public health statistics.

Acknowledgments. The results presented here are in part based upon data generated by Medipath, Cypath and the TCGA Research Network: https://www.cancer.gov/tcga. Authors would like to thank Alexandre Filiot and Auriane Riou for their valuable help and insights to conduct the downstream evaluations, and Lucas Fidon for his feedback.

Disclosure of Interests. Persons affiliated with Owkin own stock-options in the company (A.P., J.K., G.R., A.O.).

References

1. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Vaidya, A., Le, L.P., Gerber, G., Sahai, S., Williams, W., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (Mar 2024)
2. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Mahmood, F.: Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering **5**(6), 493–497 (Jun 2021)
3. Conde-Sousa, E., Vale, J., Feng, M., Xu, K., Wang, Y., Della Mea, V., La Barbera, D., Montahaei, E., Baghshah, M., Turzynski, A., Gildenblat, J., Klaiman, E., Hong, Y., Aresta, G., Araújo, T., Aguiar, P., Eloy, C., Polónia, A.: Herohe challenge: Predicting her2 status in breast cancer from hematoxylin–eosin whole-slide imaging. Journal of Imaging **8**(8) (2022)
4. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., Girard, N., Elemento, O., Nicholson, A.G., Blay, J.Y., Galateau-Sallé, F., Wainrib, G., Clozel, T.: Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature Medicine **25**(10), 1519–1525 (Oct 2019)
5. Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G.: Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. arXiv preprint arXiv:1802.02212 (2018)
6. Dawid, A.P.: The Well-Calibrated Bayesian. Journal of the American Statistical Association **77**(379), 605–610 (Sep 1982)

7. Echle, A., Ghaffari Laleh, N., Quirke, P., Grabsch, H., Muti, H., Saldanha, O., Brockmoeller, S., Van Den Brandt, P., Hutchins, G., Richman, S., Horisberger, K., Galata, C., Ebert, M., Eckardt, M., Boutros, M., Horst, D., Reissfelder, C., Alwers, E., Brinker, T., Langer, R., Jenniskens, J., Offermans, K., Mueller, W., Gray, R., Gruber, S., Greenson, J., Rennert, G., Bonner, J., Schmolze, D., Chang-Claude, J., Brenner, H., Trautwein, C., Boor, P., Jaeger, D., Gaisa, N., Hoffmeister, M., West, N., Kather, J.: Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application. ESMO Open **7**(2), 100400 (Apr 2022)
8. Filiot, A., Jacob, P., Mac Kain, A., Saillard, C.: Phikon-v2, A large and public feature extractor for biomarker prediction (2024). https://doi.org/10.48550/ARXIV.2409.09173
9. 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 (Mar 2024). https://doi.org/10.1016/j.compmedimag.2024.102337
10. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. CoRR **abs/2007.01434** (2020), https://arxiv.org/abs/2007.01434
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017)
12. Jarkman, S., Karlberg, M., Pocevičiūtė, M., Bodén, A., Bándi, P., Litjens, G., Lundström, C., Treanor, D., Van Der Laak, J.: Generalization of Deep Learning in Digital Pathology: Experience in Breast Cancer Metastasis Detection. Cancers **14**(21), 5424 (Nov 2022)
13. Kleppe, A.: Area under the curve may hide poor generalisation to external datasets. ESMO Open **7**(2), 100429 (Apr 2022)
14. Krithiga, R., Geetha, P.: Breast Cancer Detection, Segmentation and Classification on Histopathology Images Analysis: A Systematic Review. Archives of Computational Methods in Engineering **28**(4), 2607–2619 (Jun 2021)
15. Liu, Y., Kohlberger, T., Norouzi, M., Dahl, G.E., Smith, J.L., Mohtashamian, A., Olson, N., Peng, L.H., Hipp, J.D., Stumpe, M.C.: Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. Archives of Pathology & Laboratory Medicine **143**(7), 859–868 (Jul 2019)
16. McCann, R.J.: Exact solutions to the transportation problem on the line. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences **455**(1984), 1341–1380 (Apr 1999)
17. Network, T.C.G.A.: Comprehensive molecular portraits of human breast tumours. Nature **490**, 61–70 (2012), https://doi.org/10.1038/nature11412
18. Roschewitz, M., Khara, G., Yearsley, J., Sharma, N., James, J.J., Ambrózay, E., Heroux, A., Kecskemethy, P., Rijken, T., Glocker, B.: Automatic correction of performance drift under acquisition shift in medical image classification. Nature Communications **14**(1), 6608 (Oct 2023)
19. Saillard, C., Dubois, R., Tchita, O., Loiseau, N., Garcia, T., Adriansen, A., Carpentier, S., Reyre, J., Enea, D., Von Loga, K., Kamoun, A., Rossat, S., Wiscart, C., Sefta, M., Auffret, M., Guillou, L., Fouillet, A., Kather, J.N., Svrcek, M.: Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. Nature Communications **14**(1), 6695 (Nov 2023)

20. Stacke, K., Eilertsen, G., Unger, J., Lundstrom, C.: Measuring Domain Shift for Deep Learning in Histopathology. IEEE Journal of Biomedical and Health Informatics **25**(2), 325–336 (Feb 2021)
21. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., Van Der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis **58**, 101544 (Dec 2019)
22. Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, Z., Liu, J., Jin, M.: Predicting Axillary Lymph Node Metastasis in Early Breast Cancer Using Deep Learning on Primary Tumor Biopsy Slides. Frontiers in Oncology **11**, 759007 (Oct 2021)
23. Xu, X., Hou, R., Zhao, W., Teng, H., Sun, J., Zhao, J.: A Weak Supervision-based Framework for Automatic Lung Cancer Classification on Whole Slide Image. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 1372–1375. IEEE, Montreal, QC, Canada (Jul 2020)
24. Xu, Z., Verma, A., Naveed, U., Bakhoum, S.F., Khosravi, P., Elemento, O.: Deep learning predicts chromosomal instability from histopathology images. iScience **24**(5), 102394 (May 2021)
25. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-Guided Foundation Model Adaptation for Pathological Image Classification. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, vol. 14224, pp. 272–282 (2023)