Attention-based prediction of lymph node status in pt1 CRC using a histopathology foundational model

Nil Arenós Bach[1†], Debora Gil[1†], Pau Cano[1†], Eva Musulén[2†], Pau Cano[1†]

[1]Computer Vision Center, Cerdanyola del Vallès, Barcelona, Spain.
[2]Josep Carreras Institute, Badalona, Barcelona, Spain.

[†]These authors contributed equally to this work.

Abstract

Colorectal cancer is the second leading cause of cancer death in the world. If the adenocarcinoma reaches the lymph nodes, it can spread and increase the risk of metastasis. Visual biomarkers that are indicative of lymph node status could be extracted and learnt using deep learning (DL). From a multi-center cohort of 393 patients, we trained and validated a DL pipeline, using 2048x2048 px patches from haematoxylin and eosin-stained whole-slide images (WSI). For each patch, a CLS token was extracted using the UNI2 foundational model, and aggregated at patient level using an attention mechanism. With focus on clinical translation, we analysed different tissue regions and attention mechanisms. The resulting models obtained better performances than health organization guidelines and indicated that (1) current histological parameters are suboptimal and (2) tumor area is not the most important tissue in predicting lymph node status. Applying foundational models to capture more specialized information allows deep learning models to learn more meaningful information and combined with an attention mechanism, showed that several parts of the WSI, and not only the tumor, are relevant for risk prediction.

Keywords: colon cancer, attention, WSI, histopathology, lymph node status

1 Introduction

Colorectal cancer is the third most common cancer in the world and the second leading cause of cancer death [1]. While a healthy lifestyle and avoiding risk factors reduce incidence and mortality [2], early detection remains one of the best methodologies of increasing chance of survival [3]. The first stage is early-stage CRC adenocarcinoma (pT1 CRC), where the adenocarcinoma infiltrates the submucosa. If the cancerous tissue penetrates further, it can reach the lymph nodes, and risk spreading to farther areas, requiring en bloc resection. To determine whether en bloc resection is necessary, different health organizations have developed histological criteria, including poorly differentiated tumor; mucinous or presenting either lymph invasion or affected vertical margin as hazardous factors [4–6]. Using these criteria, 60-70% of patients are classified as high risk and in need of surgery [7], but only 2-10.5% show lymph node affectation [8], indicating histopathological parameters to be suboptimal and not related to lymph node metastasis (LNM). While surgery decreases the chance of recurrence and lymphatic invasion, it also entails a potential increase in risk of morbidity, discomfort related to surgery aftereffects and economic costs related to treatment [9].

The lack of predictive power of current parameters highlights the need to find new biomarkers that can complement or replace the current ones. The advances achieved in the field of computer vision in the last decades are well suited to histopathological analysis and have already been proven in haematoxylin and eosin-stain(H&E) whole slide images (WSI) for anomaly detection, tissue segmentation and classification [10]. Up to our knowledge, few works focus on LNM specifically at early-stage cancer, pT1, when clinical decisions are crucial and where cohorts can be small and multicentric [11–13]. While comparison between works is difficult due to varying datasets, inclusion parameters and methodology, existing works [14–16] have varying performances with AUC between 0.53 and 0.79.

A main challenge in the prediction of LNM is that decisions are taken at patient level, while analysis is at WSI level. Since WSI are huge images that cannot be processed in their entirety, predictions are generally based on Multiple Instance Learning (MIL) strategies. In a MIL approach, WSI are tilled into smaller area patches that are embedded by encoder models and then, aggregated to obtain a prediction at patient level, commonly with an Attention-Based mechanism (AB-MIL). Due to the costs associated with this process, generally only areas of interest are tilled, such as poorly differentiated clusters [16] or tumour tissue [15, 16]. However, some works focusing on either the entire WSI or tissue surrounding the tumour [17] have provided new insights or obtained significant improved results.

While existing AB-MIL works [11–13] for LNM prediction address valuable technical issues, they have less focus on considerations relevant for clinical translation. With this in mind, we contribute to LNM prediction models in: 1) ablation study of different attention mechanisms for MIL in small multicentric cohorts; 2) analysis of different tissue types that may complement tumor-specific features; 3) clinical relevance by benchmarking against guidelines for surgical selection; 4) analysis of attention distribution across tissue types and success in predictions.

2 Methodology

2.1 Dataset

A cohort of 393 (278 N0 and 115 N1+) patients[1] was assembled from data from 28 hospitals across all Spain. Each patient had 1-5 H&E slides from the biopsy, stained using the standard hematoxylin and eosin-stain, with Ventana and Roche technology, but with the addition of the D2-40 immunostaining to detect lymphatic vessels. Scanning was performed with Ultra-Fast 180 slide scanner at 400x magnification to obtain large, high-resolution slides. Images were inspected and curated to mitigate center-specific artifacts. Each curated slide downsampled at 1/64 of the original resolution was annotated using OpenSlide by one pathologist, to identify regions with different tissue affectation. These annotations were combined with tissue masks to obtain a semantic segmentation of tissue samples.

For each slide, non-overlapping patches of 2048x2048 pixels at 0.24 μm/pixel were extracted. While there is color variation between WSI coming from different centers, the colors also contain meaningful information and to avoid losing that information, no color correction was performed.

Clinical and pathological information from each patient was also collected, including the pathological indicators used for decision-rule guidelines: lymphovascular invasion (LI), tumor budding (TB), degree of differentiation (DG) and vertical margin affectation (VRM). For each patient, a label was established based on the number of affected nodes: 0 affected nodes (N0) or 1 or more affected nodes (N1+).

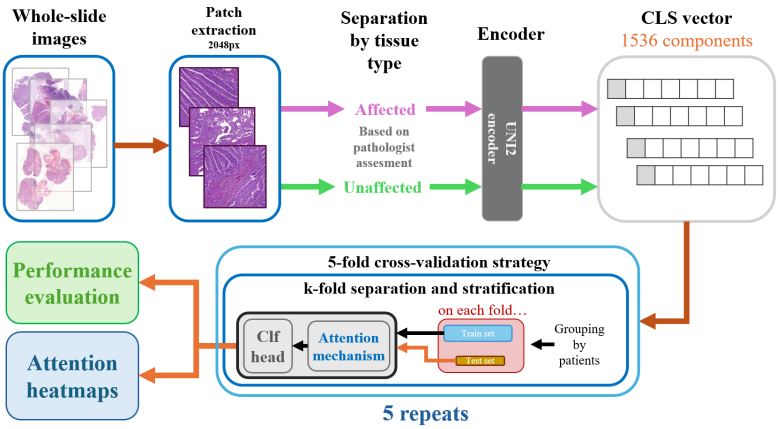

Fig. 1 Extraction and prediction pipeline for WSIs

[1] **Ethics statement.** All experiments were conducted in accordance with the Declaration of Helsinki and the International Ethical Guidelines for Biomedical Research Involving Human Subjects by the Council for International Organizations of Medical Sciences (CIOMS). Each collection of patient samples was ethically approved by its institution or health center. All authors reviewed and approved the final manuscript.

2.2 AB-MIL Strategies for LNM prediction

Fig. 1 illustrates the main steps of the AB-MIL system for LNM prediction. Patches extracted from specific tissue regions are embedded using a DL foundational model. The collection of all embeddings are the input to an attention mechanism that learns a global context of the tissue which, in turn, is the input to a binary classification head for the prediction of lymph node status.

Each slide was tiled into 2048x2048 pixels non-overlaping patches at 0.24 μm/pixel. These patches were the input to UNI2 encoder [18]. Given that UNI2 is built on the ViT-H/14 architecture via DINOv2 [18], patches were resized to 224x224 px and normalized by a 3 channel mean of 0.485, 0.456, 0.406 and a standard deviation of 0.229, 0.224, 0.225. The specific size of 2048x was selected to emulate clinical bird-eye methodology that looks at entire regions of tissue and in conjunction with the chosen foundational model, which is trained with WSI of 20mag magnification. Since our WSIs had 40mag magnification, resizing from 2048x to 224x makes our data comparable to UNI2 training images.

The 1536 dimensional vector of UNI2 encoder is aggregated for all patient patches using an attention mechanism with 4 heads, as good compromise between modelling different tissue patterns while minimizing overfitting. We considered three different attention mechanisms for MIL: a simple attention mechanism (SIMPL), a gated attention mechanism (GATED) and a self-attention mechanism (SELF). The SIMPL mechanism had a MLP of two hidden layers with 768 neurons and separated by a sigmoid activation function; GATED uses two separate linear layers of 768 neurons and 4 neurons (one for each head), with sigmoid and tanh activation respectively, followed by element wise multiplication and another hidden layer of four neurons (similar to those designed by Ilse et al. [19]. Finally, in the SELF mechanism, three vectors (Q, K, V) are linearly projected from the input features, with the same dimensions (1536), followed by the standard self-attention mechanism [20]. All three are followed by a softmax to obtain weights that are then applied to the patches.

This way, the model can incorporate information from different slides at the same time and thanks to the multi-head setup, can focus on different elements and structures.

Three patch collections were considered:

- All patches (ALL): all patches cropped from slides, regardless of tissue type.
- Unaffected area (UNAF): only patches with less than 5% affected tissue.
- Tumor area (AFF): only patches with more than 95% affected tissue.

The downsampling used to produce masks of affected and unaffected regions, motivated using 5/95% thresholds to label predominantly unaffected and tumor patches.

Pathological examinations are usually based on visual inspection of tumorous tissue. In order to assess whether predictions also focus on these type of tissue, we analysed whether there were any differences in attention distribution between AFF and UNAF areas. For each patch, the maximum from the 4 attention heads was taken to have a single attention value. A one-tailed t-student test for the average difference in maximal attentions was used to detect significance.

3 Experiments

We evaluated the impact of the aggregation step on 3 different experiments: 1) comparison of tissue region and attention mechanisms; 2) comparison with clinical guidelines and 3) assessment of attention distribution between tissue regions.

Independent neural networks were trained for each patch collection on all available H&E slides, as long as the minimum of 10 patches could be obtained. Models were trained for 30 epochs with learning rate 1e-4 and no weight decay using a weighted cross-entropy loss with weights (N0: 0.29, N1+:0.71) to mitigate class unbalance. In the first experiment, we also included a comparison to a benchmark SoA method that codifies the entire WSI. Given that we might have multiple slides for some patients, slide embeddings are also aggregated using the 3 attention mechanisms for patient-level predictions.

To mitigate leakage, data was split at the patient level and stratified by diagnosis, so the partitions had a similar data distribution. To estimate uncertainty, a 5 times repeated 5-fold cross-validation at patient level was performed, providing five different evaluations for each patient and a total of 25 folds. Metrics on each fold were aggregated to compute descriptive statistics and 95% confidence intervals. We measured performance with sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). For the analysis of attention distribution, each trial was analyzed independently. To explore biases, we performed independent analysis for correct and wrong predictions.

3.1 Results

Tissue Attention Mechanisms. Table 1 reports averages with its corresponding 95% CIs for each configuration and metric. SELF performs worse regardless of tissue type, with predictions biased towards N0. SIMPL and GATED attention perform similarly across all considered patch collections and the difference in sensitivity is not significant (p-value= 0.1801). The inclusion of all patches outperforms including only either unaffected or tumor region (p-value for sensitivity <0.05), the latter being considered in the literature as the most relevant tissue for predicting LNM. Regarding baseline SoA models aggregating WSI embeddings, they all perform similarly across attention mechanisms with lower sensitivity than SIMPL-ALL and GATED-ALL (p-values all <0.05).

Table 1 Performance metrics on the 25 folds, with accompanying CI intervals.

Selected tissue	Specificity	Sensitivity	NPV	PPV
SIMPL-WSI	0.89 (0.86, 0.92)	0.46 (0.40, 0.52)	0.80 (0.78, 0.82)	0.65 (0.60, 0.70)
SIMPL-ALL	0.87 (0.85, 0.90)	0.64 (0.60, 0.68)	0.85 (0.84, 0.87)	0.70 (0.65, 0.74)
SIMPL-UNAF	0.81 (0.78, 0.84)	0.51 (0.46, 0.56)	0.80 (0.78, 0.82)	0.52 (0.48, 0.56)
SIMPL-AFF	0.88 (0.86, 0.90)	0.55 (0.51, 0.59)	0.83 (0.81, 0.84)	0.66 (0.61, 0.70)
GATED-WSI	0.87 (0.83, 0.91)	0.47 (0.39, 0.55)	0.80 (0.78, 0.82)	0.63 (0.58, 0.68)
GATED-ALL	0.88 (0.86, 0.90)	0.60 (0.55, 0.65)	0.84 (0.83, 0.86)	0.68 (0.64, 0.72)
GATED-UNAF	0.82 (0.79, 0.85)	0.51 (0.47, 0.56)	0.80 (0.79, 0.82)	0.55 (0.51, 0.59)
GATED-AFF	0.87 (0.84, 0.90)	0.54 (0.49, 0.60)	0.82 (0.81, 0.84)	0.65 (0.61, 0.70)
SELF-WSI	0.84 (0.81, 0.87)	0.48 (0.43, 0.53)	0.80 (0.79, 0.81)	0.57 (0.52, 0.62)
SELF-ALL	0.91 (0.82, 1.00)	0.28 (0.09, 0.47)	0.74 (0.64, 0.84)	0.34 (0.12, 0.56)
SELF-UNAF	0.80 (0.66, 0.94)	0.10 (0.00, 0.22)	0.72 (0.58, 0.86)	0.12 (0.00, 0.25)
SELF-AFF	0.90 (0.78, 1.00)	0.34 (0.15, 0.53)	0.80 (0.67, 0.93)	0.36 (0.16, 0.56)

Comparison to Clinical Guidelines. We chose 3 referent health organization guidelines: the European Society for Medical Oncology (ESMO), the National Comprehensive Cancer Network (NCCN) and Japanese Society for Cancer of the Colon and Rectum (JSCCR). Table 2 shows equal metrics than Table 1 for these guidelines. The same repeated k-fold methodology was performed as with the models, providing CI intervals. All guidelines have higher specificity than sensitivity, with ESMO and JSCCR with higher sensitivity (0.62) at the cost of a drop in specificity (0.7 in comparison to NCCN 0.75). Comparing to MIL models in Table 1, SIMPL-ALL and GATED-ALL have similar sensitivity with improved specificity.

Table 2 Performance metrics of the guidelines, with CI intervals.

Guidelines	Specificity	Sensitivity	NPV	PPV
ESMO	0.70 (0.68, 0.72)	0.62 (0.58, 0.66)	0.82 (0.81, 0.84)	0.46 (0.44, 0.48)
NCCN	0.75 (0.73, 0.77)	0.54 (0.50, 0.57)	0.80 (0.78, 0.81)	0.47 (0.45, 0.49)
JSCCR	0.70 (0.68, 0.72)	0.62 (0.58, 0.66)	0.82 (0.81, 0.84)	0.46 (0.44, 0.48)

Attention Distribution. Table 3 shows the mean attention in the two tissue areas (AFF and UNAF), and the p-value of the t-test for correct predictions and failures for the SIMPL-ALL model. For the correctly classified patients, the null hypothesis is only refuted for the N1+ patients, where the affected tissue is the most important for LNM prediction. In the N0 patients, the unaffected tissue receives higher attention. For the incorrectly classified patients, the opposite is observed, showing that the model is looking at the incorrect tissue type.

Figure 2 shows the attention distribution across the five repeats for a correctly classified N1+ case. We observe that the highest attention patches are found in two regions across all repeats: in the top half of the affected tissue and on the bottom-right

region of the unaffected area, with the rest of the tissue being considered generally unimportant.

Table 3 Attention Distribution Comparison. Boostrapping mean and CI are provided as e-4.

	N0			N1+		
Repeat	AFF	UNAF	p-value	AFF	UNAF	p-value
Correctly classified patients						
1	8.96	15.79	1.000	13.95	10.87	<0.01
2	9.15	16.22	1.000	13.86	10.15	<0.01
3	8.68	15.58	1.000	12.76	10.54	<0.01
4	8.55	17.44	1.000	14.18	10.66	<0.01
5	9.35	15.14	1.000	14.51	9.71	<0.01
Incorrectly classified patients						
1	12.87	15.79	0.999	16.68	22.83	1.000
2	17.18	10.63	<0.01	18.56	20.34	0.999
3	14.24	11.13	<0.01	16.47	18.60	0.999
4	14.74	12.42	<0.01	16.58	26.58	1.000
5	17.64	11.34	<0.01	17.32	20.42	0.999

4 Discussion

Comparison between AB-MIL and clinical guidelines shows that AB-MIL outperforms all three guidelines in both sensitivity and specificity. This indicates that attention might be incorporating to predictions visual elements not considered in current guidelines. Comparison between SIMPL-ALL and GATED-ALL shows not enough significant difference in sensitivity (p-value= 0.1801) to establish distinctiveness between the two mechanisms, but SIMPL-ALL significantly outperforms both SELF-ALL (p-value <0.001) and its baseline, SIMPL-WSI (p-value <0.001).

Both performance analysis and statistical tests indicate that the entire WSI is relevant to prediction. The affected tissue by itself had slightly better performance than unaffected, showing that the tumour remains a key component of LNM prediction, but that the rest of the tissue needs to be considered. Visual observation of the attention heatmaps of a correctly predicted N1+ patient (Fig. 2) illustrates attention is distributed between affected and unaffected tissue in specific spots. Our statistical analysis of attention distribution in successful predictions also indicates that models use features of both tissue types and with higher weight on unaffected regions for N0, contrary to literature that only indicates focusing on affected tissue. Further experiments should try to discover what specific areas are important, even incorporating other histological data or parameters, such as genetic biomarkers [17], which reported good performances in other studies.

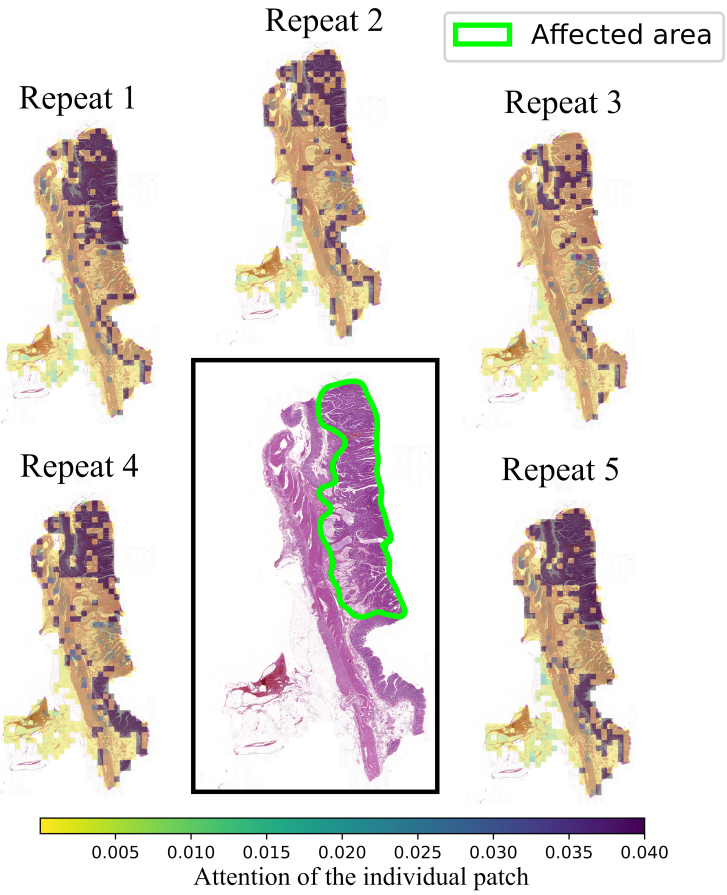

Fig. 2 Attention heatmap of correctly classified N1+ patient. In green, the tumorous tissue. In the center, the original WSI, marked in black. On the corners, the attention maps corresponding to the five repeats, with patches colored based on their attention weight.

Limitations. First, our study was limited by its retrospective nature and its repeat methodology, which contains intrinsic dependencies that cannot be solved and would need an external dataset to validate the results. Second, our dataset presents certain limitations regarding its scale and distribution. With 393 patients spread across 28 different centers, the small number of cases per site may introduce center-specific signatures, potentially leading to shortcut learning [21]. The difference in light, technician manipulation, and biopsies introduce variability, which combined with pathologist WSI curation minimize these signatures. We also checked that LNM distribution was similar, and observed no performance bias between centers. Finally, our analysis of AB-MIL mechanisms is based on UNI2 feature extractor. Other patch-level encoder models (like the ones available at the TRIDENT library developed by the Mahmood Lab [22]) should be also analysed to check that the findings on tissue type relevance are not biased by the patch encoder model.

5 Conclusions

Our model can correctly separate the two classes using a simple 4-head attention mechanism, while incorporating information from the entire slide. Incorporating a histopathology foundational model allowed feature extraction be specifically designed to the problem at hand and is partly responsible for the performance metrics. The attention mechanism seems to focus on different areas from traditional methodology and we validated the results with a statistical test. Additionally, we also report that current health guidelines are suboptimal and new parameters need to be devised to complement or replace the current ones. While we understand the methodology presents training-test dependencies that we cannot correct for, we believe our findings are promising and merit further exploration.

References

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, *et al.*: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)

[2] Logan, R.F., Patnick, J., Nickerson, C., Coleman, L., Rutter, M.D., Wagner, C.: Outcomes of the bowel cancer screening programme (bcsp) in england after the first 1 million tests. Gut **61**(10), 1439–1446 (2012)

[3] Seeff, L.C., Royalty, J., Helsel, W.E., Kammerer, W.G., Boehm, J.E., Dwyer, *et al.*: Clinical outcomes from the cdc's colorectal cancer screening demonstration program. Cancer **119**, 2820–2833 (2013)

[4] Benson, A.B., Venook, A.P., Adam, M., Chang, G., Chen, Y.-J., Ciombor, et al.: Colon cancer, version 3.2024, nccn clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network **22**(2D) (2024)

[5] Cervantes, A., Adam, R., Roselló, S., Arnold, *et al.*: Metastatic colorectal cancer: Esmo clinical practice guideline for diagnosis, treatment and follow-up. Annals of Oncology **34**(1), 10–32 (2023)

[6] Hashiguchi, Y., Muro, K., Saito, Y., Ito, Y., Ajioka, Y., Hamaguchi, *et al.*: Japanese society for cancer of the colon and rectum (jsccr) guidelines 2019 for the treatment of colorectal cancer. International journal of clinical oncology **25**(1), 1–42 (2020)

[7] Choi, Y.S., Kim, W.S., Hwang, S.W., Park, S.H., Yang, *et al.*: Clinical outcomes of submucosal colorectal cancer diagnosed after endoscopic resection: a focus on the need for surgery. Intestinal research **18**(1), 96–106 (2020)

[8] Backes, Y., Bergeijk, J., Borg, F., Schwartz, *et al.*: Risk for incomplete resection after macroscopic radical endoscopic resection of t1 colorectal cancer: a multicenter cohort study. Official journal of the American College of Gastroenterology—ACG **112**(5), 785–796 (2017)

[9] Zaffalon, D., Daca-Alvarez, M., Gordoa, K., Pellisé, M.: Dilemmas in the clinical management of pt1 colorectal cancer. Cancers **15**(13), 3511 (2023)

[10] Jabbar, M.A.D.A., Yim, W.W., Wee, F., Chong, L.Y., Yeong, J.: 1250 H&E-based cell prediction multi-classification models to capture morphologically distinct subpopulations of CD8+ T cells. BMJ Publishing Group Ltd (2024)

[11] Feng, M., Zhao, Y., Chen, *et al.*: A deep learning model for lymph node metastasis prediction based on digital histopathological images of primary endometrial cancer. Quantitative Imaging in Medicine and Surgery **13**(3), 1899 (2023)

[12] Gao, F., Jiang, L., Guo, T., Lin, J., Xu, W., Yuan, L., Han, Y., Yang, J., Pan, Q., Chen, E., *et al.*: Deep learning-based pathological prediction of lymph node metastasis for patient with renal cell carcinoma from primary whole slide images. Journal of Translational Medicine **22**(1), 568 (2024)

[13] Xu, F., Zhu, C., Tang, W., Wang, Y., Zhang, Y., Li, J., Jiang, H., Shi, *et al.*: Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. Frontiers in oncology **11**, 759007 (2021)

[14] Kwak, M.S., Lee, H.H., Yang, J.M., Cha, *et al.*: Deep convolutional neural network-based lymph node metastasis prediction for colon cancer using histopathological images. Frontiers in Oncology **10**, 619803 (2021)

[15] Muti, H.S., Röcken, C., Behrens, H.-M., Löffler, C.M., Reitsam, N.G., Grosser, B., Märkl, *et al.*: Deep learning trained on lymph node status predicts outcome from gastric cancer histopathology: a retrospective multicentric study. European Journal of Cancer **194**, 113335 (2023)

[16] Pai, R.K., Hartman, D., Schaeffer, D.F., Rosty, C., Shivji, S., Kirsch, R., Pai, R.K.: Development and initial validation of a deep learning algorithm to quantify histological features in colorectal carcinoma including tumour budding/poorly differentiated clusters. Histopathology **79**(3), 391–405 (2021)

[17] Mo, S., Zhou, Z., Dai, W., Xiang, W., Han, L., Zhang, L., Wang, R., Cai, S., Li, Q., Cai, G.: Development and external validation of a predictive scoring system associated with metastasis of t1-2 colorectal tumors to lymph nodes. Clinical and translational medicine **10**(1), 275–287 (2020)

[18] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, *et al.*: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)

[19] Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136 (2018). Pmlr

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[21] Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, o.: Distilling foundation models for robust and efficient models in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 162–172 (2025). Springer

[22] Zhang, A., Jaume, G., Vaidya, A., et al.: Accelerating data processing and benchmarking of ai models for pathology. arXiv preprint arXiv:2502.06750 (2025)