A rule-based approach as an alternative to end-to-end TIL prediction in gastroesophageal adenocarcinoma

David R Wessels[*1] D.R.WESSELS@UVA.NL
Ylva A Weeda[*2,3] Y.A.WEEDA@AMSTERDAMUMC.NL
Hoel Kervadec[1,2] H.T.G.KERVADEC@UVA.NL
Sybren L Meijer[2,3] S.L.MEIJER@AMSTERDAMUMC.NL
Erik J Bekkers[1] E.J.BEKKERS@UVA.NL
[1] *University of Amsterdam, Amsterdam, The Netherlands*
[2] *Amsterdam UMC, Amsterdam, The Netherlands*
[3] *Cancer Center Amsterdam, Cancer Treatment and Quality of Life, Amsterdam, The Netherlands*

Editors: Under Review for MIDL 2025

Abstract

Tumour-Infiltrating Lymphocytes (TILs) are a critical biomarker for predicting immunotherapy response in gastroesophageal adenocarcinoma (GEA). However, manual quantification of TILs on Whole Slide Images (WSIs) is labour-intensive and suffers from high inter-observer variability. While Deep Learning offers a solution for consistent assessment, current approaches typically fall into two distinct categories: interpretable, multi-step rule-based pipelines that mimic clinical workflows, or end-to-end Multiple Instance Learning (MIL) models that operate as "black boxes." These paradigms are rarely evaluated side-by-side. In this study, we propose a novel rule-based TIL quantification pipeline, comprising tissue segmentation and cell detection steps, and benchmark it against a standard end-to-end MIL approach. We utilize the AUMC-SELECT-AI dataset, a multi-center cohort combining TCGA and Amsterdam UMC data, to train our component models using dense ROI annotations. We evaluate both approaches on a held-out test set of 100 WSIs with pathologist-derived TIL scores. By comparing these methods, we assess the trade-offs between the granular interpretability of rule-based systems and the data-driven flexibility of end-to-end learning, providing insight into their clinical applicability for GEA biomarker assessment.

Keywords: Pathology, Computer Vision, Rule-Based,

1. Introduction

Pathology plays a pivotal role in cancer care, providing the foundation for diagnosis and biomarker assessment, both of which offer prognostic insights and guide treatment selection. Haematoxylin and eosin (H&E) staining is standard clinical practice in histopathology, representing detailed morphological information about both the tumour and its microenvironment. While these tissue images are exceptionally information-rich, their size and complexity makes consistent interpretation challenging. The assessment of tumour-infiltrating

[*] Contributed equally

lymphocytes (TILs) provides a clear example. During TIL assessment, pathologists must evaluate the proportion of tumour-associated stroma occupied by mononuclear cells (i.e. lymphocytes and plasma cells). Accurately quantifying TIL density across large tissue regions is inherently challenging, as it exceeds what can reliably be achieved by the naked eye. As a result, current manual TIL assessment is highly susceptible to intra- and interobserver variability [1] [2] [3]. Computational methods offer a promising solution for TIL assessment, as they can systematically quantify cells across entire slides when digitized into high-resolution whole slide images (WSIs). Such methods can provide reproducible, consistent measurements, overcoming the limitations of manual evaluation of WSI. Rule-based methods and end-to-end deep learning (DL) models are both widely applied in computational pathology. However, the two approaches are rarely evaluated side-by-side, despite representing fundamentally different trade-offs in terms of performance, flexibility, interpretability and robustness. In this study, we propose a new rule-based TIL-detection pipeline and benchmark it against an end-to-end multiple instance learning (MIL) model. We evaluate their performance and medical applicability on standard pathology H&E-stained WSIs of gastroesophageal adenocarcinoma (GEA). By comparing both approaches in terms of performance, interpretability, and generalizability, we aim to provide a clear assessment of how an interpretable pipeline relates to a fully learned MIL approach for capturing meaningful immune-tumour interactions.

2. Related works

2.1. TILs as biomarker for immunotherapy response

The introduction of immune checkpoint inhibitors (ICIs), a form of immunotherapy, has transformed the landscape of cancer treatment by harnessing the patient's own immune system to attack cancer cells. Nonetheless, a substantial proportion of patients fail to respond favourably, leading to futile cancer treatment and increased economic burden [4], [5]. There is an urgent need for reliable prognostic biomarkers to identify those patients most likely to benefit from ICIs. Tumour infiltrating lymphocytes (TILs) are a key component in a successful anti-cancer immune response and are quantified as the percentage of tumour-associated stromal area occupied by mononuclear immune cells (i.e. lymphocytes and plasma cells). Both TIL density and their spatial distribution have shown to be strong predictors for treatment efficacy in breast cancer [6], lung cancer [7] and melanoma [8]. TILs can be assessed on H&E-stained slides using several visual steps, as proposed by International TILs Working Group guidelines [9]. Pathologists typically define the tumour and stromal areas, identify immune cell types, and estimate the percentage of TILs. Additional considerations, such as which regions to include or exclude, further complicate scoring. Figure 1 illustrates the main workflow to convey the core idea of how a TIL score is determined in a gastroesophageal adenocarcinoma (GEA) tissue sample.

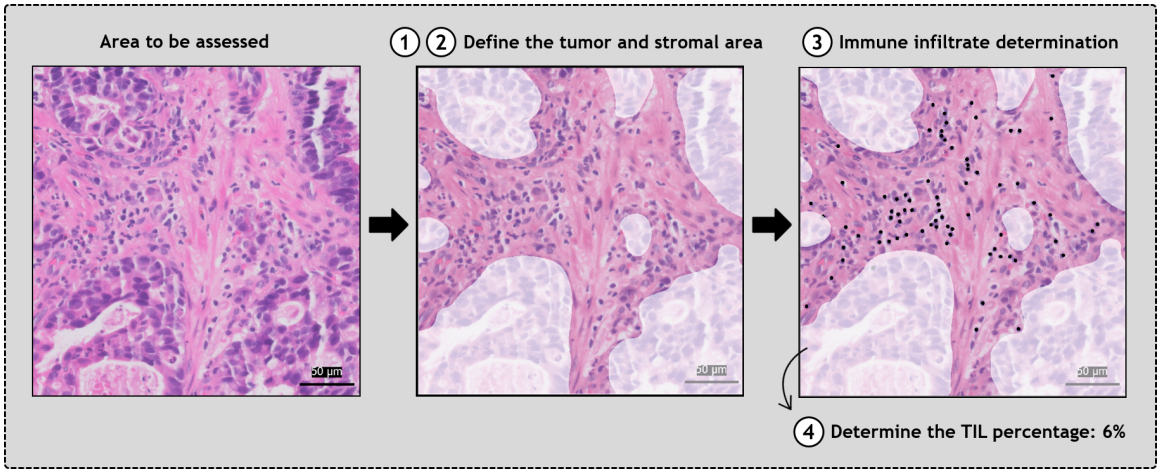

Figure 1: Proposed framework for TIL assessment in gastroesophageal carcinoma specimens, following these key steps 1) Define the tumour area, 2) Define the stromal area, 3) Immune infiltrate determination and 4) Determine the TIL percentage. The steps are illustrated using a selected region from a whole slide image of a patient with oesophageal adenocarcinoma.

2.1.1. Previous Work on TIL quantification

Current computational approaches for TIL quantification can be broadly grouped into three categories: patch-based classifiers, slide-level MIL models, and rule-based or cell-level segmentation methods.

Patch-based models classify fixed-size tiles as TIL-positive or TIL-negative using either convolutional neural networks (CNNs) [10] or classical machine-learning approaches [11]. Although these models operate in an end-to-end manner at the patch level, they provide only coarse spatial information and lack whole-slide contextual reasoning. Typically, patch-level predictions are aggregated heuristically to produce slide-level scores or heat-maps.

Slide-level multiple instance learning (MIL) models [12] consider the entire WSI as a bag of patches, learning to predict a single score directly from slide-level labels. Attention maps generated by these models can highlight which regions are most "important" for the prediction, but recent studies indicate that these maps often lack reliability and may not correspond to biologically meaningful areas [13] [14]. Region-aware attention models [15] partially address this issue by explicitly modelling tissue compartments and maintaining patch-level spatial information. Despite these advances, most MIL models still output binary results (TIL-high vs. TIL-low). While binary classification may be useful for triage or cohort stratification, such outputs are of limited practical value for clinical-decision making, as optimal TIL thresholds across cancer types have not yet been established.

Cell-level and rule-based pipelines [16] [17] focus on nuclear segmentation and lymphocyte identification, providing detailed information about the location or distribution of TILs within the tissue. While these methods count lymphocytes across large tumour regions or predefined regions of interest, they do not distinguish between stromal and intratumoural compartments, which is essential for clinically compliant TIL reporting. Building on these principles, Thagaard et al. [18] developed a workflow that explicitly incorporates

key elements of the manual TIL-assessment guidelines into an automated rule-based model. Although this represents an important step toward clinically aligned quantification, the approach relies heavily on labour-intensive manual cell-level annotations. Nevertheless, when such annotations yield models with high accuracy and consistent clinical performance, this investment can be well justified.

Despite the growing interest in both rule-based pipelines and end-to-end MIL frameworks for TIL detection, their respective tradeoffs have not been systematically evaluated. To our knowledge, no studies have directly compared rule-based cell-level pipelines with end-to-end MIL models for TIL detection on the same dataset, particularly in GEA. This motivates our work, in which we benchmark a rule-based TIL detection workflow against an end-to-end MIL model to evaluate their relative strengths and medical applicability.

3. Method

Our proposed rule-based pipeline automates the clinical estimation of TILs by mimicking the pathologist's workflow. The pipeline, illustrated in Figure 2, consists of three sequential stages: tissue segmentation, cell detection, and TIL quantification.

First, a U-Net model segments the WSI into distinct tissue compartments. We specifically target the tumour-associated stroma, which serves as the region of interest (ROI) for TIL assessment. The segmentation model classifies tissue into 'tumour', 'stroma', 'inflamed stroma' and 'background' categories. For the final calculation, we define the stromal area as the union of the 'stroma' and 'inflamed stroma' classes, excluding the tumour nests themselves.

To ensure that the analysis focuses on the main tumour mass and avoids irrelevant tissue areas, we generate a 'tumor bulk' mask. This is derived from the tumour segmentation mask by applying a sequence of morphological operations: binary dilation, downscaling, Gaussian blurring, thresholding, and upscaling. This process effectively merges fragmented tumour clusters into a cohesive region representing the tumour bed, while filtering out isolated tumour cells or small clusters that may be artifacts or non-representative.

Parallel to segmentation, we employ the CellViT++[19] model to detect individual immune cells across the entire slide. We specifically filter for mononuclear immune infiltrate (lymphocytes and plasma cells). Each detected lymphocyte is treated as a circular object with a fixed radius of 8 μm.

Finally, the TIL score is computed as an area-based ratio, consistent with international guidelines:

$$\text{TILs} = \frac{\sum \text{Area}(\text{Lymphocytes} \in \text{Stroma})}{\text{Area}(\text{Stroma})} \times 100 \tag{1}$$

The total lymphocyte area is the sum of the areas of the individual fixed-radius circles ($A = \pi r^2$). The stromal area is computed as the number of stromal pixels multiplied by the pixel area (in squared microns). Note that the exact units cancel out in the division. This approach ensures that the score reflects the density of the immune infiltrate specifically within the relevant stromal compartment, ignoring intratumoral lymphocytes or those in necrotic regions.

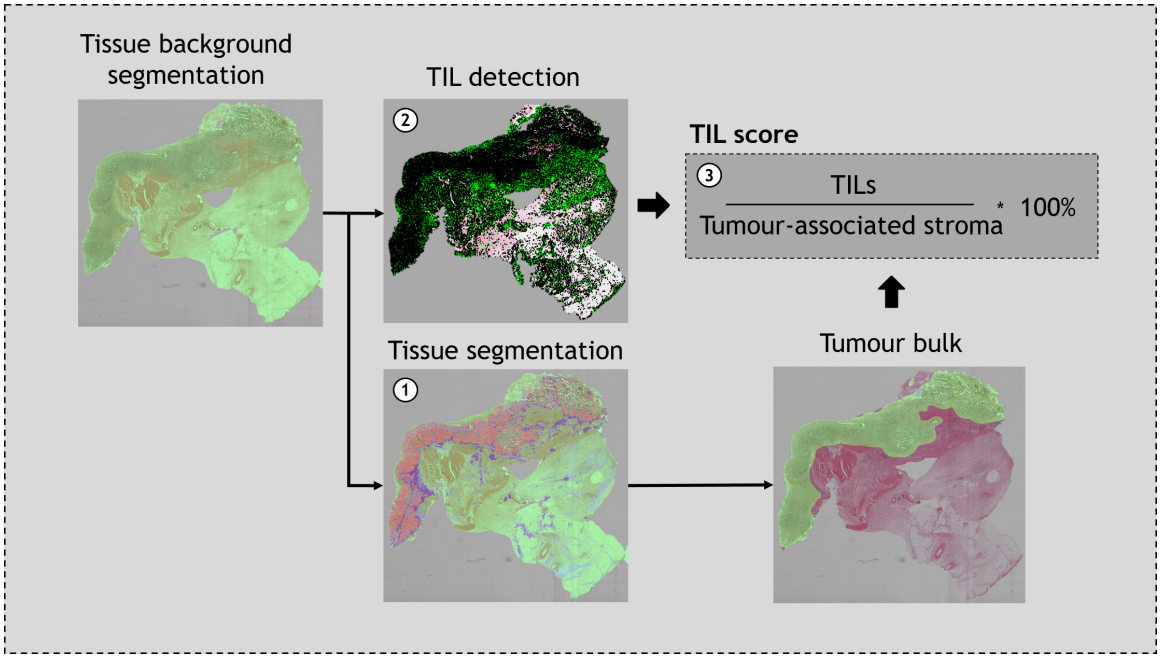

Figure 2: Schematic representation of the rule-based TIL detection pipeline. The process involves (1) segmenting the tissue to identify the stromal compartment, (2) detecting lymphocytes using CellViT++, and (3) calculating the TIL score as the ratio of lymphocyte area to stromal area.

4. Experiments

4.1. Datasets

The dataset (AUMC-SELECT-AI) used in this study consists of digitized WSIs of H&E-stained pretreatment biopsies and untreated resection specimens (no preoperative chemotherapy/and or radiation) of GEA patients. These WSIs were obtained from the public The Cancer Genome Atlas Program (TCGA) dataset, where the microscopic slides were digitized using various microscopic scanners (e.g., Leica GT450/AT2/CS2 and Hamamatsu S60/360). Additional samples were retrieved from GEA patients treated at Amsterdam University Medical Center. These patients were included in various consortia between 1989 and 2023 and provided consent for reuse of their data. For the Amsterdam UMC cohorts, the slides were digitized with a Philips IntelliSite Ultra-fast Scanner. All WSIs were scanned at a resolution of 0.25 µm/pixel and checked for adequate image quality (e.g. high-resolution, sharp-focused). To construct the training and test sets, WSIs were randomly sampled from the combined dataset while preserving the relative proportions of TCGA and Amsterdam UMC samples in both subsets, thereby avoiding cohort-based sampling bias.

Training set: The training set contains 525 regions of interest across 93 WSIs, which were manually annotated by medically trained PhD students for thirteen tissue classes (e.g., tumor, stroma, muscle, fat, necrosis) and checked by expert pathologists. The regions of interest had an average size of approximately 500 µm. In addition, 109 regions of interest across 30 additional WSIs were annotated for eight specific cell types (e.g., lymphocytes,

plasma cells, tumour cells, epithelial cells), with region sizes ranging from 100-200 µm. These cell annotations yielded a total of ˜31.000 individual cell annotations.

Test set: The held-out test set comprises 100 WSIs that were assessed by two expert pathologists following the International TILs Working Group guidelines, resulting in slide-level TIL scores for each case.

4.2. Evaluation of sub-components TIL-pipeline

To segment the tissue into clinically relevant compartments, we trained a U-Net architecture [20] using the dense ROI annotations described in Section 4.1. The UNet model was trained on the 525 regions with segmentation annotation discussed in 4.1, for 300 epochs using the Adam optimizer with standard hyperparameters. While the dataset contains thirteen tissue classes, for the purpose of TIL quantification, we aggregated these into three primary categories: tumour, stroma, and ignore. As explained in the method, this segmentation step defines the region of interest for the subsequent cell detection. After training the model acquired a validation intersection-over-union (IoU) of 0.7052 and a pixel accuracy of 0.8622. Moreover, in figure 3 we show training and validation curves of the loss.

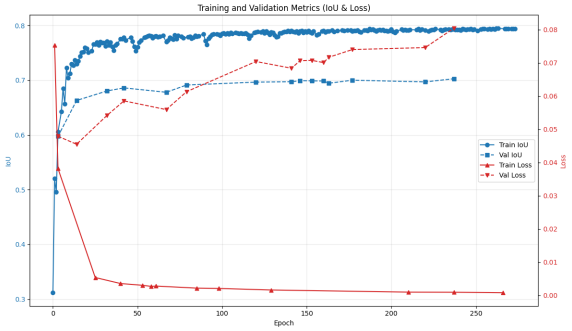

Figure 3: Training and validation metrics for the UNet segmentation model. The plots illustrate the progression of loss and Intersection over Union (IoU) over 300 epochs.

For cell detection, we employed CellViT++ [19], a state-of-the-art vision transformer-based model. We evaluated two variants of this model. First, we used the off-the-shelf CellViT++ pretrained on the PanNuke dataset, which detects five standard cell classes: neoplastic, inflammatory, connective, dead, and epithelial. Second, to improve specificity for our GEA cohort, we finetuned the CellViT++ classifier head using our cell-level annotations. This finetuned model was trained to discriminate between five specific cell types relevant to our analysis: lymphocytes, granulocytes, plasma cells, tumour cells, and other. This allowed for a more granular assessment of the immune infiltrate compared to the generic "inflammatory" class of the standard model.

A representative five-fold cross-validation evaluation of the fine-tuned CellViT++ model is shown in Table 1. The results demonstrate that the model can reliably distinguish the major immune cell subtypes from tumour cells across different data splits. Although overall performance is high, the scores display some variation between folds, most notably for lymphocytes and plasma cells, which indicates that discrimination between these immune

populations is more challenging and slightly less consistent. This suggests that while the fine-tuned model provides a more detailed characterisation of the immune infiltrate than the off-the-shelf version, there remains some variability in cell-level performance across training splits.

Table 1: Five-fold cross-validation performance (F1 scores) of the finetuned CellViT++ model. The table presents the classification scores for each of the five target cell types, illustrating the model's ability to discriminate between immune cell subtypes and tumor cells across different data splits.

Cell Type	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Lymphocytes	0.83	0.86	0.80	**0.45**	**0.35**
Granulocytes	0.96	0.59	0.81	0.77	0.73
Plasma cells	0.82	0.65	0.63	**0.41**	**0.20**
Tumor cells	0.72	0.78	0.83	0.97	0.87
Other	0.83	0.80	0.81	0.65	0.80
Mean	**0.78**	**0.74**	**0.77**	**0.61**	**0.57**

4.3. Correlations with a pathologist

To evaluate the clinical relevance of our proposed methods, we compared the automated TIL scores against the gold-standard scores provided by expert pathologists on the held-out test set of 100 WSIs.

Rule-based approach: As described in Section 3.1, the rule-based pipeline calculates the TIL density as the ratio of the area occupied by detected lymphocytes to the total area of tumour-associated stroma. We evaluated this pipeline using both the off-the-shelf CellViT++ and our finetuned version.

MIL Baseline: As a comparison, we trained a standard Attention-based Multiple Instance Learning (ABMIL) model [21]. This end-to-end approach aggregates patch-level features extracted using a CTransPath encoder [22] to predict a slide-level TIL score directly from the WSI. The model was trained using the pathologist-derived TIL scores as weak supervision.

To ensure a fair comparison with the rule-based pipeline, which was evaluated on the held-out test set of 100 WSIs, we employed a Leave-One-Out Cross-Validation (LOOCV) strategy for the MIL baseline. In this procedure, the model is trained N times (where $N = 100$), each time using $N - 1$ slides for training and the single remaining slide for evaluation. Since the MIL model requires slide-level labels for training, directly training it on the same 100 WSIs used for testing the rule-based model would constitute data leakage. Conversely, splitting this limited set into smaller train/test splits would result in an underpowered model and unstable evaluation. By using LOOCV, we maximize the training data for each fold while generating an unbiased prediction for every slide in the dataset. This allows us to benchmark the performance of the end-to-end approach on the exact same set of slides as the rule-based method, without overfitting to the test set.

Table 2 presents the Pearson correlation coefficients between the automated scores and the pathologist's assessment. We report correlations for both the continuous TIL scores ('Pearson Real') and for clinically relevant binned scores ('Pearson Buckets').

As shown in Table 2, the MIL baseline exhibits substantially lower performance compared to our rule-based pipeline, with Pearson correlations of 0.26 and 0.12 for the real and bucketed TIL scores, respectively. This outcome is expected, as the MIL model was trained using leave-one-out cross-validation on a small held-out set, providing limited exposure to the data and preventing the model from reaching its full potential.

In contrast, the rule-based CellViT++ pipeline demonstrates high correlations with manual pathology assessment even with the off-the-shelf model (0.79 and 0.76), reflecting the benefits of extensive training with cell-level and tissue-level annotations. Fine-tuning the CellViT++ classifier to distinguish lymphocytes and plasma cells further improves performance by focusing on immune cell types relevant to tumour elimination, in line with manual TIL scoring guidelines. This targeted approach likely accounts for the additional gain in correlation.

Overall, these results highlight the strengths of annotation-driven, rule-based pipelines in capturing clinically meaningful TIL patterns, while also illustrating the challenges of weakly supervised MIL approaches in data-limited settings.

Table 2: Pearson Correlation over 100 WSIs comparing TILs pipeline with the end-to-end approach.

Model	**Pearson Real**	**Pearson Buckets**
MIL	0.2596	0.1242
Rule-Based (all inflammatory cells)	0.7909	0.7579
Rule-Based (only plasma cells & lymphocytes)	0.8021	0.7632

5. Conclusion

The rule-based pipeline enables accurate, clinically aligned TIL quantification and closely mirrors pathologist reasoning. It outperforms the weakly supervised MIL baseline on our dataset, though the comparison primarily highlights the impact of extensive annotation rather than a definitive performance gap. This side-by-side evaluation illustrates the complementary strengths of interpretable, annotation-driven pipelines versus end-to-end models, while leaving open the potential for larger or better-trained MIL models to achieve stronger generalisation in future work.

Acknowledgments

This research was supported by a public grant from the 'Koningin Wilhelmina Fonds' (KWF150591) The funder had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript or in the decision to publish the results.

References

[1] Carsten Denkert et al. "Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group". In: *Modern Pathology* 29 (2016), pp. 1155–1164. DOI: 10.1038/modpathol.2016.109.

[2] M. R. Van Bockstal et al. "Interobserver variability in the assessment of stromal tumor-infiltrating lymphocytes (sTILs) in triple-negative invasive breast carcinoma influences the association with pathological complete response: the IVITA study". In: *Modern Pathology* 34 (2021), pp. 2130–2140. DOI: 10.1038/s41379-021-00865-z.

[3] N. B. Baharun et al. "Assessment of Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer: Interobserver Variability and Contributing Factors". In: *Diagnostics* 15 (2025). DOI: 10.3390/diagnostics15192492.

[4] S. Das and D. B. Johnson. "Immune-related adverse events and anti-tumor efficacy of immune checkpoint inhibitors". In: *Journal for ImmunoTherapy of Cancer* 7 (2019), p. 306. DOI: 10.1186/s40425-019-0805-8.

[5] N. Schaft et al. "The future of affordable cancer immunotherapy". In: *Frontiers in Immunology* 14 (2023), p. 1248867. DOI: 10.3389/fimmu.2023.1248867.

[6] P. Schmid et al. "Pembrolizumab plus chemotherapy as neoadjuvant treatment of high-risk, early-stage triple-negative breast cancer: results from the phase 1b open-label, multicohort KEYNOTE-173 study". In: *Annals of Oncology* 31 (2020), pp. 569–581. DOI: 10.1016/j.annonc.2020.01.072.

[7] J. D. Fumet et al. "Prognostic and predictive role of CD8 and PD-L1 determination in lung tumor tissue of patients under anti-PD-1 therapy". In: *British Journal of Cancer* 119.8 (2018), pp. 950–960. DOI: 10.1038/s41416-018-0220-9.

[8] A. I. Daud et al. "Tumor immune profiling predicts response to anti–PD-1 therapy in human melanoma". In: *Journal of Clinical Investigation* 126.9 (2016), pp. 3447–3452. DOI: 10.1172/JCI87324.

[9] R. Salgado et al. "The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014". In: *Annals of Oncology* 26.2 (2015), pp. 259–271. DOI: 10.1093/annonc/mdu450.

[10] H. Zhang, L. Chen, L. Li, et al. "Prediction and analysis of tumor infiltrating lymphocytes across 28 cancers by TILScout using deep learning". In: *npj Precision Oncology* 9 (2025), p. 76. DOI: 10.1038/s41698-025-00866-0.

[11] T. B. Fisher et al. "Digital image analysis and machine learning-assisted prediction of neoadjuvant chemotherapy response in triple-negative breast cancer". In: *Breast Cancer Research* 26.1 (2024), p. 12. DOI: 10.1186/s13058-023-01752-y.

[12] Yoni Schirris et al. *WeakSTIL: Weak whole-slide image level stromal tumor infiltrating lymphocyte scores are all you need*. 2021. arXiv: 2109.05892 [eess.IV]. URL: https://arxiv.org/abs/2109.05892.

[13] Minjae Chung et al. "Evaluating Visual Explanations of Attention Maps for Transformer-Based Medical Imaging". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops*. Springer Nature Switzerland, 2025, pp. 110–120. ISBN: 9783031776106. DOI: 10.1007/978-3-031-77610-6_11. URL: http://dx.doi.org/10.1007/978-3-031-77610-6_11.

[14] Willem Bonnaffé et al. *Beyond attention: deriving biologically interpretable insights from weakly-supervised multiple-instance learning models*. 2023. arXiv: 2309.03925 [q-bio.QM]. URL: https://arxiv.org/abs/2309.03925.

[15] R. Perera, P. Savas, D. Senanayake, et al. "Annotation-efficient deep learning for breast cancer whole-slide image classification using tumour infiltrating lymphocytes and slide-level labels". In: *Communications Engineering* 3 (2024), p. 104. DOI: 10.1038/s44172-024-00246-9.

[16] A. Liu et al. "Prognostic Significance of Tumor-Infiltrating Lymphocytes Determined Using LinkNet on Colorectal Cancer Pathology Images". In: *JCO Precision Oncology* 7 (2023), e2200522. DOI: 10.1200/PO.22.00522.

[17] M. Chou, I. Illa-Bochaca, B. Minxi, et al. "Optimization of an automated tumor-infiltrating lymphocyte algorithm for improved prognostication in primary melanoma". In: *Modern Pathology* 34 (2021), pp. 562–571. DOI: 10.1038/s41379-020-00686-6.

[18] J. Thagaard et al. "Automated Quantification of sTIL Density with H&E-Based Digital Image Analysis Has Prognostic Potential in Triple-Negative Breast Cancers". In: *Cancers* 13.12 (2021), p. 3050. DOI: 10.3390/cancers13123050.

[19] Fabian Hörst et al. "Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models". In: *arXiv preprint arXiv:2501.05269* (2025).

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer. 2015, pp. 234–241.

[21] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[22] Xiyue Wang et al. "Transformer-based unsupervised contrastive learning for histopathological image classification". In: *Medical image analysis* 81 (2022), p. 102559.