Aligning Computational Pathology with Clinical Practice: A Systematic Review of AI Tools for Pathology Report Elements in Colorectal Cancer

Elias Baumann[1,2] *, José F Carreño-Martinez[1,2] *, Ana Leni Frei[1,2] *, Javier Garcia-Baroja[1,2] *, Mauro Gwerder[1,2] *, Amjad Khan[1] *, Rina Mehmeti[1,2] *, Jacob Hanimann[1,2], Philipp Zens[1], Heather Dawson[1], Alessandro Lugli[1], Inti Zlobec[1]

1 Institute of Tissue Medicine and Pathology, University of Bern, Bern, Switzerland

2 Graduate School for Cellular and Biomedical Sciences (GCB), University of Bern, Bern, Switzerland

*Authors have contributed equally to this work and are sorted alphabetically

Abstract:

BACKGROUND: Pathology reporting of colorectal cancer (CRC) follows the International Collaboration on Cancer Reporting (ICCR) guidelines which define a set of 25 elements (such as tumor grade, TNM stage and microsatellite instability) to be assessed for diagnosis. With the aim to further develop the CRC diagnostic routine, multiple computational tools have been developed in the last ten years. Despite the excellent sensitivity and potential advantages, including reduced inter-observer variability, many tools do not reach clinical deployment. This suggests that there are critical challenges to address when developing these algorithms given the aim to reliably and automatically assess ICCR elements.

METHODS: To summarize existing efforts in deep- and machine learning for ICCR CRC elements and highlight existing gaps between development and clinical deployment, this systematic review collected studies on computational tools for colorectal cancer histopathology analysis published between 2015 and 2024.

RESULTS: In total, 4863 studies were retrieved for the analysis, of which 66 remained after screening. Most of the reviewed studies focus on a subset of just three ICCR elements, namely mismatch repair status, BRAFV600E mutation testing, and lymph

node status. Notably, many of the studies did not include clinically relevant and validated results, which puts into question their reliability for routine diagnosis.

CONCLUSION: These results show the wide gap between research and clinical practice in pathology with the example of CRC diagnosis. There is an unmet need for publicly available datasets addressing a variety of topics, and a stronger focus on clinically important tasks. This review is critical to help the community align their contributions in computation pathology with the clinic and ultimately increase the translational potential of the developed tools.

1. Introduction

Digital pathology is shaping the future of the fast-evolving field of pathology and digitalized pathology laboratories are an increasing trend worldwide[1]. Once these laboratories have the digital infrastructure in place (e.g. slide scanner and viewer, laboratory information systems or image management system etc.), they can benefit from AI-based tools such as AI algorithms in supporting pathologists to analyze tissue slides. Leveraging AI is potentially helpful for time-consuming cases that require assessment of long lists of criteria and are prone to interobserver variability. Moreover, several studies have shown increased diagnosis confidence among pathologists when assisted with AI[2–4].

AI-based tools have been developed in numerous different pathologies, including colorectal cancer (CRC). CRC is the third most common cancer worldwide with a raising incidence in younger ages due to sedentary lifestyle and consumption of processed food [5]. As a result of this, and widespread screening, CRC constitutes a considerable portion of cases in a pathologist's diagnostic routine.

On one hand, the World Health Organization (WHO) Classification covers all generally accepted criteria[6] for standardized disease definition and classification. On the other hand, the College of American Pathologists (CAP) and ICCR have developed evidence-based guidelines for standardizing reporting in pathology that are implemented

globally[7,8]. In this review, we focused on the ICCR CRC guideline[7] which covers both the macroscopic as well as microscopic evaluation of CRC resection specimen. As automatic methods have been predominantly developed for microscopic assessment, we restrict the review to only include reporting elements from this category. These elements are classified into “core” elements that are key for disease management, staging and prognosis, such as histological tumor type and grade, lymphatic invasion, TNM staging; and "non-core” elements, recommended but not validated by all levels of evidence, including tumor budding, mismatch repair (MMR) status, coexistent pathology, etc. While various AI algorithms have been developed to assess individual core and non-core elements, it remains unclear whether they can be combined for an AI-based evaluation of all elements on the digitized whole slide image (WSI) and produce an automated CRC pathology report. With this in mind, through this systematic review we aimed to reflect on:

1. The current landscape of deep learning (DL) in computational pathology (CPath) for CRC
2. Published work on the 25 elements of the ICCR CRC diagnostic report
3. Gaps towards automated ICCR-based reporting
4. The potential avenues and future steps to close these gaps

2. Methods

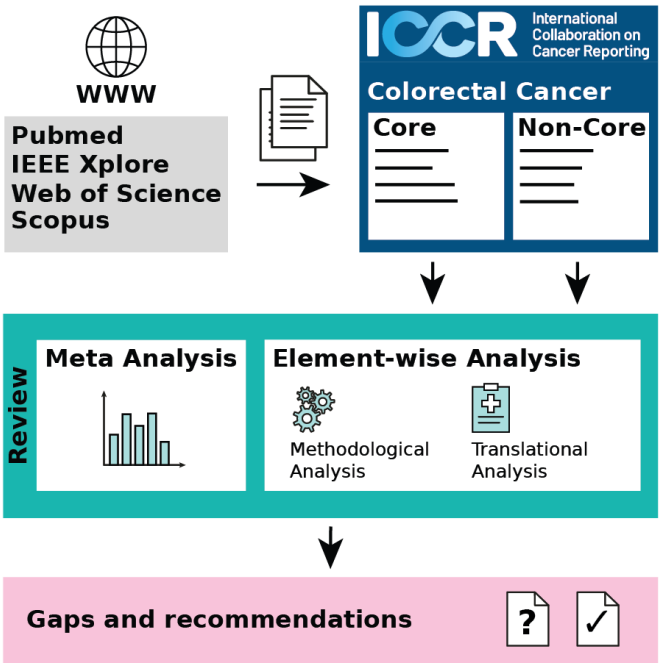

Figure 1: Graphical overview of the review.

2.1 **Structure of this review**

Given the breadth of this review, the Results section is divided into three major parts, starting with general meta-results on gathered papers, and followed by meta-analysis of each element of the ICCR CRC guideline grouped here as core or non-core. The elements are sorted by number of included publications. After this analysis, and a general discussion on the most important takeaways, we include recommendations for elements where currently no readily usable computational approach has been published (Fig. 1).

Review Protocol

We perform a systematic literature review by following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines[9].

Search strategy

For the review, we defined the following search criteria, with slight adaptations specific to the search engine. Essentially, search terms included a term related to cancer, to colon/colorectal, and a term related to histopathology to focus on the specific disease. Furthermore, deep learning, artificial intelligence, computer vision, and computer

assistance-related search terms were added to cover all papers that develop computer-aided approaches. To include a wide range of publications across disciplines in our study, we included Pubmed, Web of Science, Embase, and IEEE Xplore as databases. Only papers published between January 1st 2015 and September 13th 2024 were eligible for inclusion in this study. Only abstracts in English language were included. All article abstracts were screened in covidence[10]. The full search terms for each database can be found in Supplementary Table 1.

Article Screening and eligibility

Articles were screened by at least two raters, with a third rater reviewing the abstract in case of disagreement. Duplicates were removed automatically by covidence[10], or manually, if the automatic method failed to recognize the duplicate. Initially, abstracts were screened to assess relevance for the review according to a list of exclusion criteria. Specifically, articles should include an evaluation on human CRC, and results on histopathology image data. Articles that did not include original research (such as reviews, letters and commentaries) as well as conference abstracts were also excluded. Moreover, multimodal studies were excluded if they did not include results using exclusively histopathology/clinical data. If the approach evaluated non-hematoxylin and eosin (H&E) images, did not report slide-level or patient-level results, or only worked in a semi-automatic fashion they were also excluded. Finally, the article had to be related to or directly address one of the elements, e.g. if a study concerned itself with tissue type segmentation but is not usable for or applicable to one of the elements, it was excluded. Moreover, the developed model had to produce an output closely related to how the respective element is specified in the ICCR CRC guideline. In a second round, full texts are assessed for eligibility. The same exclusion criteria were applied. The full list of exclusion criteria can be found in Supplementary Table 2.

Adding references from other sources

Additionally, some articles from other sources highly relevant to this review were included. These papers were identified via references or by using search engines such as google scholar or semantic scholar for the specific ICCR CRC guideline element. Moreover, some articles did not contain the necessary information of the search criteria

in their abstract, but were still relevant to this review, so they were manually included as well.

Data extraction

To thoroughly assess individual studies, information was extracted from each study. Specifically, data concerning the cohorts used (including staining, number of patients, cancer stage distribution, target distribution), the training setup (dataset splits, cross validation and involvement of pathologists), the DL pipeline (input size, pre-processing, model, post-processing, training task and target) and reported results (AUC, F1, ... if available). Additionally, we assessed for each study whether there were slide-level or patient-level results, if the published model is clinically approved, whether data, code, or model weights were available and whether there is any information on inference time. Data were extracted by a single reviewer and verified by a second reviewer. Many studies report multiple results across several cohorts and in multiple configurations. To properly represent those studies in our analysis, the best results were extracted from each paper unless specified otherwise in the specific chapter. For multimodal models which use data that would normally not be used in a pathology report, we report the results where only histopathology data is used. Papers which cover more than one ICCR element will be included multiple times in the analysis but evaluated specifically for the respective topic.

3. Results

The search across databases yielded 4863 papers in total, of which 1636 were identified as duplicates. Additionally, 20 papers from citation search or additional manual search were included, yielding 3247 studies to be screened. In accordance with the exclusion criteria, 3066 studies were subsequently excluded, leaving 181 studies for full text screening. After full text screening, 66 studies were included in the final review (Figure 2).

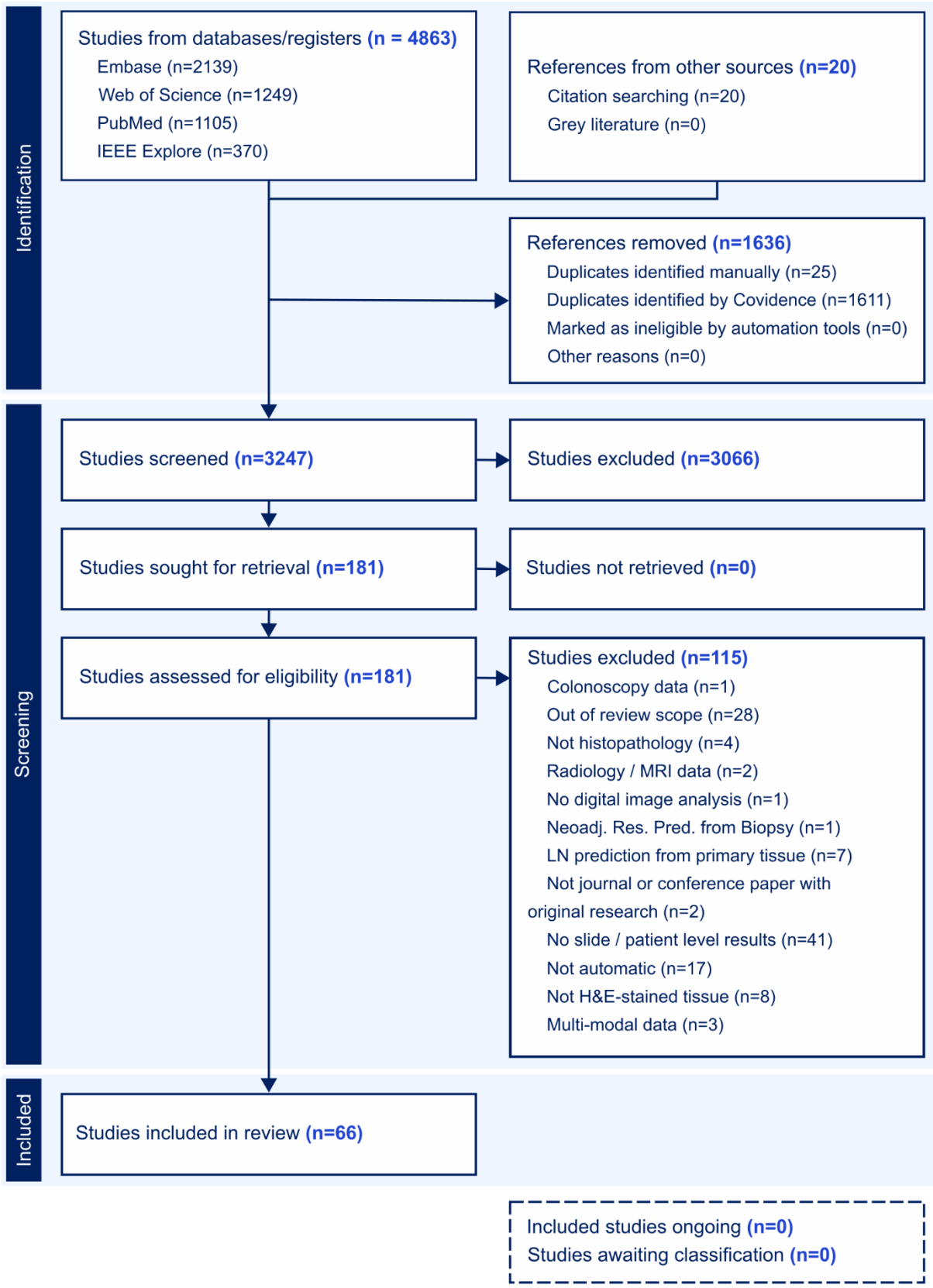

Figure 2: Flowchart following the PRISMA guidelines

3.1 Meta Results

Although half of the ICCR CRC report elements are covered by the 66 selected studies, most publications concern only three topics. Specifically, 39/66 (59%) of studies showed prediction results on MMR status, 16/66 (24%) on BRAF V600E mutation, and 9/66 on pN (14%). Coexistent pathologies (5/66), Grade (4/66), and TNM Stage (3/66) are covered by five or less publications. For tumor budding and perineural invasion two studies were included each, whereas one study each for pT and response to neoadjuvant therapy was extracted. Publications can cover multiple topics at once, for example, 14 publications cover both MMR status and BRAF V600E mutation. For the remaining elements of the ICCR CRC guidelines the employed search and exclusion criteria did not yield any publications (Figure 3 A).

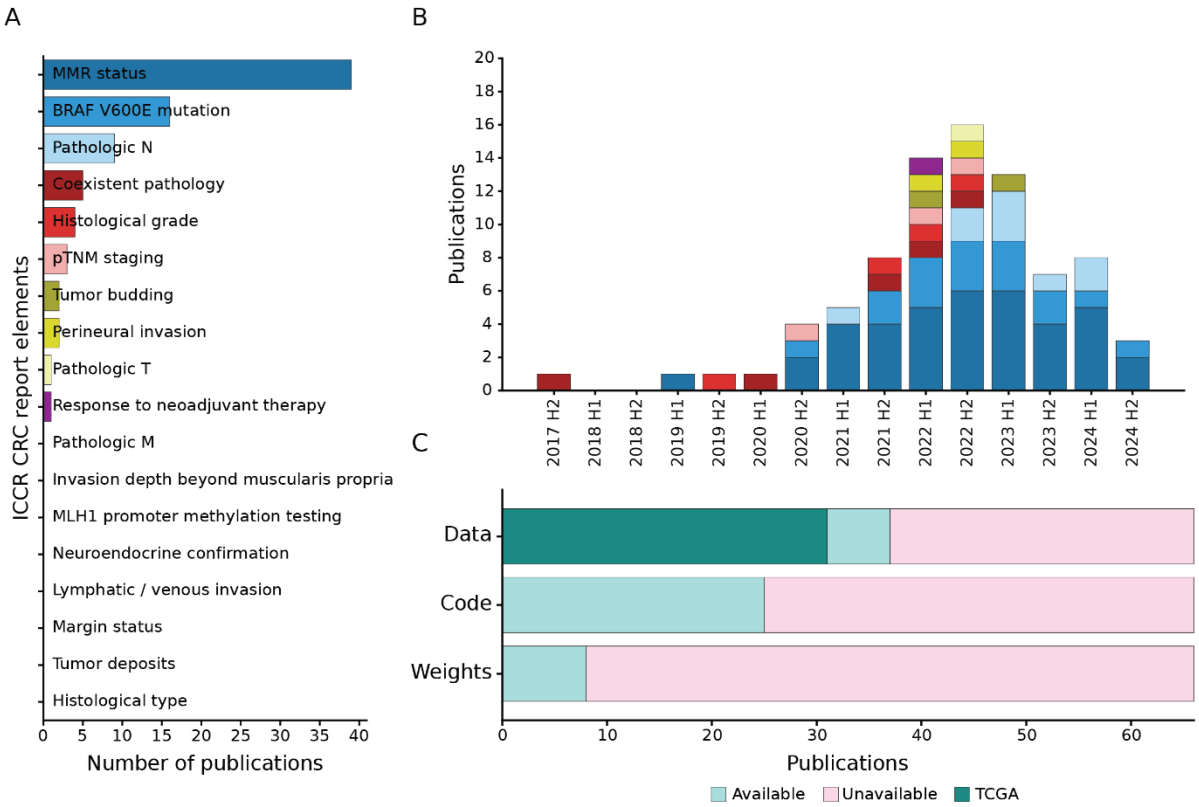

Figure 3: A) Number of articles covering a specific element of the ICCR guidelines. B) Number of published articles over time, aggregated by half-year. Studies for the second half of 2024 were only retrieved until 13.09.2024. C) Number of papers with accessible data, code and model weights respectively.

Following the general trend of increased attention and impactful contributions in CPath[11], we also observe an overall increase in publications over time, even when applying strict exclusion criteria (Figure 2 B). MSI and BRAF mutation prediction studies remain the most common over the last four years. Lymph node metastasis prediction studies were increasingly published starting in the second half of 2022. Overall, the topic variety among published studies also increased in the past five years, but without a substantial growth in the number of publications in popular topics since 2022.

Considering data, code and model weight availability (Figure 2 C), more than half of the published papers (37/66, 56%) use some publicly available dataset or make their data available, and 31/66 (47%) use The Cancer Genome Atlas (TCGA) in some form. Code was available for less publications (25/66, 38%) and publications that made model weights available were scarce (8/66, 12%). It should be noted that resources which are available "upon reasonable request" were considered unavailable, although we did not investigate whether the information could have indeed been obtained in these cases.

This initial analysis has shown substantial variation in coverage across all reporting elements. In the next step, we will analyze published work related to the core elements of the ICCR CRC guideline.

3.2 ICCR CRC Guidelines: Core Elements

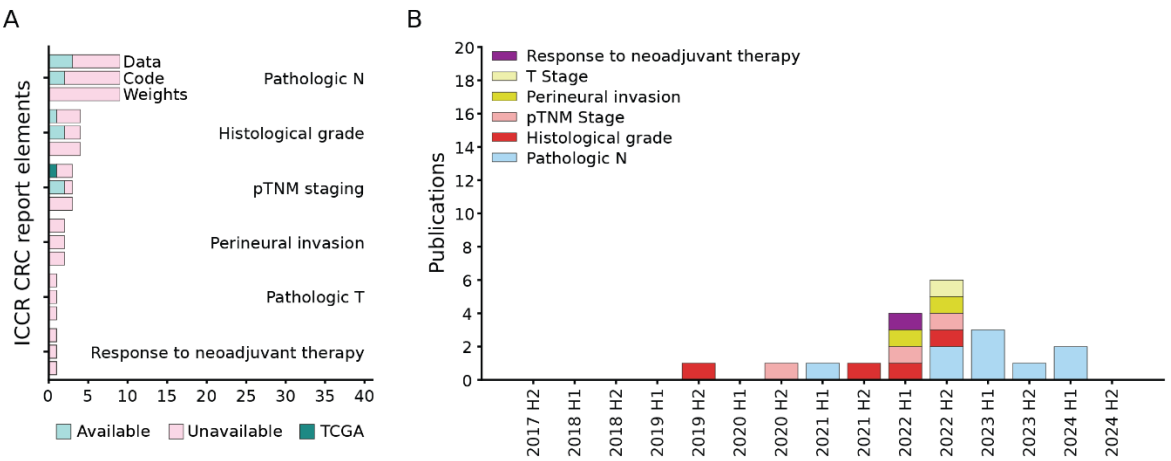

Figure 4: A) Number of articles covering a specific core element of the ICCR guidelines, split by the data, code, and weight availability for each topic. B) Number of published articles over time, aggregated by half-year. Not shown are elements where no publications passed exclusion criteria.

Elements classified as "Core" within the ICCR CRC guidelines are considered essential for clinical management and staging of CRC patients. Excluding clinical and macroscopically retrieved information, the microscopic core elements comprise histological tumor type, tumor grade, extent of invasion (and pT), lymphatic, venous and perineural invasion, lymph node status (and pN), tumor deposits, response to neoadjuvant therapy, margin status, margin status, histologically confirmed distant metastasis (and pM), the aggregation into pTNM, and confirmatory evaluation for neuroendocrine neoplasms. Of these twelve core elements, six have been addressed in 20 published articles, with pN (9/20) and tumor grade (4/20) receiving the most attention. 5/20 publications have used publicly available data or released their datasets, 6/20 made their code available, but none published model weights. Next, we present the state of research of each individual core element in order of the number of included publications.

Regional Lymph Nodes (pN) and Lymph Node Status

Regional lymph node (LN) status is crucial for determining adjuvant chemotherapy and cancer staging. The TNM classification of regional LNs (pN) includes: NX (nodes cannot be assessed), N0 (no metastasis), N1a (1 positive node), N1b (2-3 positive nodes), N1c (tumor deposits in soft tissue without nodal metastasis), N2a (4–6 positive nodes), and N2b (≥7 positive nodes)[12]. A LN is considered positive if it contains at least one metastatic lesion larger than 2 mm[12]. Pathologists aim to evaluate at least twelve nodes per case, though this number can be influenced by factors such as specimen length, patient age, or neoadjuvant therapy[7,13].

Study	Year	Validation Cohorts (WSI)	Best results	clinical validation	External Validation	Data	Code	Weights
Chuang[14]	2021	1 (1000)	0.997 AUC	-	-	-	-	-
Mainovskaya[15]	2022	1(-)	0.863 DICE	-	-	-	-	-
Tsuneki[16]	2022	1(928*)	0.911 Sens. 0.931 Spec. 0.928 ACC	-	✓	-	-	-

Bándi[17]	2023	1 (44)	0.979 AUC	-	-	✓	-	-
Khan[18]	2023	4 (172-2803)	1.000 Sens. 1.000 Spec.	✓	✓	-	-	-
Tan[19]	2023	1 (169)	0.977 AUC 0.953 Sens. 0.935 PPV 0.935 ACC	-	-	-	-	-
Kindler[20]	2023	2(74-288)	0.940 AUC 0.820 Sens. 0.980 Spec. 0.906 PPV 0.952 ACC	✓	-	-	-	-
Giammanco[21]	2024	7 (652-4911)	1.000 Sens. 0.990 Spec. 0.991 ACC	-	✓	-	-	-
Huang[22]	2024	1 (137)	0.8877 F1** 0.8024 Sens.** 0.969 ACC**	✓	-	✓	✓	-

Table 1: Extracted studies for lymph node status and pN. *187 CRC WSI, ** Pathologist performance assisted by model.

Data and Methods

Several approaches have been used to automatically screen LN slides for metastases, listed in Table XX. Most methods are focused on metastasis detection/segmentation or based on metastasis detection further predicting the slide-level labels, whether positive or negative. Model evaluation was performed on both external (30% of the studies) and internal cohorts (70% of the studies). Clinical validation or pathologist involvement was reported in ~30% of studies.

Huang et al.[22] used a human-in-the-loop AI system (nuclei.io) to personalize ML models for colorectal LN metastasis detection. The approach improved accuracy, F1 score, sensitivity, and especially detection of isolated tumor cells, while significantly reducing evaluation time for fellows and negative cases. Similarly, Kindler et al.[20] trained a Deep Neural Network tool achieving high pixel-level accuracy. Clinical testing showed high sensitivity (0.990), strong interobserver agreement ($\kappa = 0.94$), and significantly reduced review times for pathologists. Furthermore, Khan et al.[18] demonstrated 100% agreement between their ensemble model and expert pathologists in a study of 217 cases.

Clinical Employability

The reviewed studies show high sensitivity and specificity across the board for metastasis detection. However, the automation of pN staging requires the additional step of counting the number of positive lymph nodes in a slide. This can be difficult due to tissue preparation steps where some large lymph nodes may be cut in half and may be present twice in the same slide, or on multiple different slides. Correctly re-identifying the same lymph node and ensuring positive lymph nodes not being counted twice therefore remains an unsolved problem. One approach to addressing this issue could be a better link between macroscopic tissue preparation and AI-based image analysis. For example, lymph node placement information could be additionally stored, but consistent tissue inking may also be an avenue to solve the problem automatically.

Histological Tumor Grade

Tumor grade describes the level of differentiation of tumor cells where a higher grade indicates a loss of glandular organization[6,23]. In accordance with the WHO classification, the ICCR CRC guidelines classify the grade by the least differentiated component of the neoplastic lesion which then results in either low or high tumor grade.

Study	Year	Classes	Validation Cohorts (WSI)	Best Results	External Validation	Data	Code	Weights
Rathore[24]	2019	WD, MD, PD	1 (174)	0.98 AUC 0.928 Sens. 0.958 Spec. 0.95 Acc	-	✓	-	-
Schrammen[25]	2021	NORM, WD+MD, PD	1 (805)	0.751 AUC 0.877 Sens. 0.528 Spec. 0.599 Acc 0.513 F1	✓	-	✓	-
Chen[26]	2022	NORM, WD, MD, PD	1 (108)	0.981,0.983, 0.963, 0.963 AUC	-	✓	-	-
Soldatov[27]	2022	NORM, WD, MD, PD	1 (357)	0.981, 1.0, 0.933, 1.0 Spec. 0.553, 0.0, 0.577, 0.0 Sens. 0.838, 0871, 0.876, 0.974 Acc 0.696, 0.0, 0.6, 0.0 F1	-	-	-	-

Table 2: Extracted studies for tumor grade. Normal (NORM), Well-differentiated (WD), Moderately-differentiated (MD), Poorly-differentiated (PD).

Data and Methods

The retrieved studies use both private and public datasets, but with the exception of one study[25], none validate their results externally. Given the subjectivity and high interobserver variabililty of tumor grade assessment[28], a consensus label based on multiple pathologists might be more reliable as ground truth. Some studies[26,27] used two raters, with a third pathologist to resolve any conflicts. Yet, none of the retrieved studies reported agreement between pathologists and deep learning models. Most studies rely on tile classification[25–27] to directly predict tumor grade. Rathore et al.[24] use morphological and texture-based image features aggregated in a support vector machine classifier and majority voting. They achieve comparable performance to tile classifiers on their dataset. For tile classification methods, the overall strategies include multiple instance learning[25] for end-to-end training, and supervised tile-based training[26,27]. No studies aggregate at the patient level, although clinical practice typically considers the highest grade across multiple slides. This is particularly relevant for TCGA-based studies[26], as WSIs were not selected for grading and may not be representative of the overall tumor grade.

Clinical Employability

Data accessibility varies significantly across studies. While Rathore et al.[24] use the publicly available GLaS dataset and Chen et al.[26] TCGA-COAD and TCGA-READ, and Soldatov et al.[27] only make their dataset available upon request. Beyond TCGA and challenge datasets, independent validation is limited due to restricted access to additional clinical data.

Reproducibility is another concern, as most studies do not share code or model weights. Only Schrammen et al.[25] have a publicly available repository. Despite these limitations, the availability of public datasets (e.g., GLaS) enables further research and tool development.

pTNM

Pathological TNM staging is the combined assessment of pT, pN, and pM. The TNM stage directly informs treatment decisions including whether a resection is necessary, or

whether additional adjuvant therapy needs to be performed. It remains the most important prognostic factor in CRC diagnosis, with Stage I patients' five-year survival being over 90% for colon and rectum, but Stage IV 11-15% respectively [Citation Needed].

Study	Year	Validation Cohorts (WSI)	Best Results	Actionable Results	External Validation	Data	Code	Weights
Levy[29]	2021	1 (17)	0.92 AUC (Stage I,II vs III,IV)	-	-	-	✓	-
Kumar[30]	2022	2 (130-197)	0.78 - 0.81 AUC (Stage II vs IV)	✓	-	-	✓	-
Pei[31]	2022	1 (459)	0.89 Acc, 0.89 F1 (tile level)	-	-	✓	-	-

Table 3: Extracted studies for pTNM. *Results from Image-only approach, paper reports improved results when integrating genomic data.

While TNM Staging could be algorithmically constructed out of evaluated T, N, and M, all three retrieved approaches directly predict TNM stage from the primary, with only one publication also considering lymph node slides[29]. Two publications rely on TCGA COAD (and READ) for the evaluation, and Kumar et al.[30] additionally include an internal cohort for training and validation. Two methods used graph convolution to spatially integrate tissue tile embeddings into slide level predictions[29,31], whereas the last followed the more traditional method of tumor area detection, and subsequent tile classification and averaging based on morphological features of nuclei[30].

<u>Clinical Employability</u>

Only Kumar et al.[30] generate clinically actionable results, but they use their model for the specific task to differentiate between metastatic colon cancer and local colon cancer. While Pei et al.[29] also show slide-level results, they never evaluate them or report a tile-aggregation strategy. Levy et al.[29] predict low (I, II) vs high (III, IV) stage, but their approach does evaluate the slide for invasion depth so further development could lead to a comprehensive TNM analysis.

Taken together, none of the retrieved papers demonstrate an end-to-end pipeline that can reliably assess the TNM stage (I-IV) from histological slides. While there are interesting cases where directly predicting the risk of metastasis from the primary might be sensible (e.g. what is the risk of a missed metastasis), simpler approaches following

303 the UICC TNM guidelines by aggregating information about local, regional and distant
304 spread may be more clinically relevant for routine reporting.

305 **Perineural Invasion**

306 Perineural Invasion (PNI) is defined as tumor growth along a nerve, where the tumor
307 surrounds at least one-third of the nerve's perimeter and invades any of its layers[32]. PNI
308 is reported as either present or absent, and its presence has been linked to poor
309 prognosis, particularly in Stage II[33].

Study	Year	Validation cohort (WSI)	Best results	External validation	Data	Code	Weights
Han[34]	2022	1 (50)	ACC: 0.81	-	✓	-	-
Jung[35]	2022	1 (-)	AUC: 0.92	-	-	-	-

310 **Table 4:** Extracted studies for perineural invasion.

311 *Data and Methods*

312 Both included studies employ a similar approach for detecting PNI. First, nerve and
313 tumor regions are separately segmented, and subsequently a boundary between the
314 two regions is identified. While Jung et al.[35] use a private dataset for training, Han et al.[34]
315 trained on the PAIP2021 dataset comprising of 240 WSIs from multiple cancer types.
316 The dataset was used for the PAIP2021 challenge[36] to detect PNI.

317 *Clinical Employability*

318 Neither of the included studies provide the code or the weights of the DL algorithms that
319 produced the reported results. Moreover, the validation cohorts, compared to studies on
320 other topics, e.g. the previously discussed lymph node metastasis detection, are
321 comparatively small. Beyond this reproducibility gap, the methods do report a binary
322 output for the perineural invasion in accordance with the ICCR CRC guidelines.

323 **pT and Extent of invasion**

324 The extent of invasion of the primary tumor is the first part of the TNM staging and
325 reflected in pT. It is assessed by considering the different components of the colon wall
326 and the presence of tumor within them. pT ranges from pT0 (no tumor) to pT4 (tumor

327 invades other organs or perforation into the visceral peritoneum) and has a direct impact
328 on the treatment of a patient[12].

Study	Year	Classes	Validation Cohorts (Patients)	Best Results	External validation	Data	Code	Weights
Ju[37]	2022	pT 1-4	2 (38-42)	0.87, 0.90, 0.93, 0.75 AUC 0.80, 0.83, 0.88, 0.71 Sens. 0.85, 0.86, 0.88, 0.75 Spec.	✓	-	-	-

329 **Table 5:** Extracted studies for extent of invasion and pT.

330 Only one study that automates the prediction of pT in CRC passed the exclusion criteria[37].
331 In this study, the authors use two validation cohorts, one internal with 38 patients, one
332 external with 42 patients. For classification into pT1-4, they first segment the tumor area,
333 then use a patch-based classifier to predict pT for each tile to finally choose the highest
334 pT value as the final prediction. The model reaches AUCs up to 0.93 on the internal test
335 set and 0.90 on the external validation cohort.

336 Yet the validation sets only include few patients, while data, code, and weights are not
337 publicly available. They do show results with explicit cutoffs, but only in 1-vs-all settings
338 without multiclass evaluation. Therefore, more work is needed to develop robust
339 approaches for automatic pT assessment with larger validation cohorts and investigating
340 challenges in real world datasets.

341 **Response to Neoadjuvant Therapy**

342 Tumor regression gradings (TRGs) are used to assess the response to neoadjuvant
343 chemoradiotherapy (nCRT) by measuring the extent of tissue changes caused by the
344 treatment[38]. This is a common assessment step in locally advanced rectal cancer treated
345 with neoadjuvant therapies. Multiple TRGs exist but all rely on the assessment of tumor
346 to fibrosis ratio within the tumor bed[38].

Study	Year	Classes	Validation cohort (Patients)	Best Results	External validation	Data	Code	Weights
Zhang[39]	2022	Good response (TRG0-1), bad response (TRG2-3)	1 (30)	0.759 AUC 0.717 ACC	-	-	-	-

347 **Table 6:** Extracted studies for response to neoadjuvant therapy.

348

Data and Methods

We retrieved only one study that predicted regression after nCRT in a fully automated way[39]. The authors used an MIL approach with gated attention weight normalization with a final bilinear attention for multi-scale feature fusion to classify responders from non-responders according to slide-level labels. Yet, they validated the model on a breast cancer metastasis dataset rather than rectal cancer which reduces the value of the validation.

Clinical Employability

There exist many different systems for TRG, and in the included study, the AJCC system[40] was used. However, the four AJCC TRG categories were merged into a binary classification task which does not correspond to the ICCR CRC guidelines. The reviewed study trained for slide-level predictions but does not discuss aggregation strategies for patient-level predictions. However, this post-processing aggregation step is needed for successful clinical deployment, as pathologists report a patient-level TRG.

Due to the challenges caused by the TRG definitions, we did not find any study that presents a clinically ready algorithm to assess tumor regression in rectal cancer, yet a tissue type segmentation model fine-tuned to predict remaining tumor and fibrosis could replicate every TRG scheme.

3.3 ICCR CRC Guidelines: Non-Core Elements

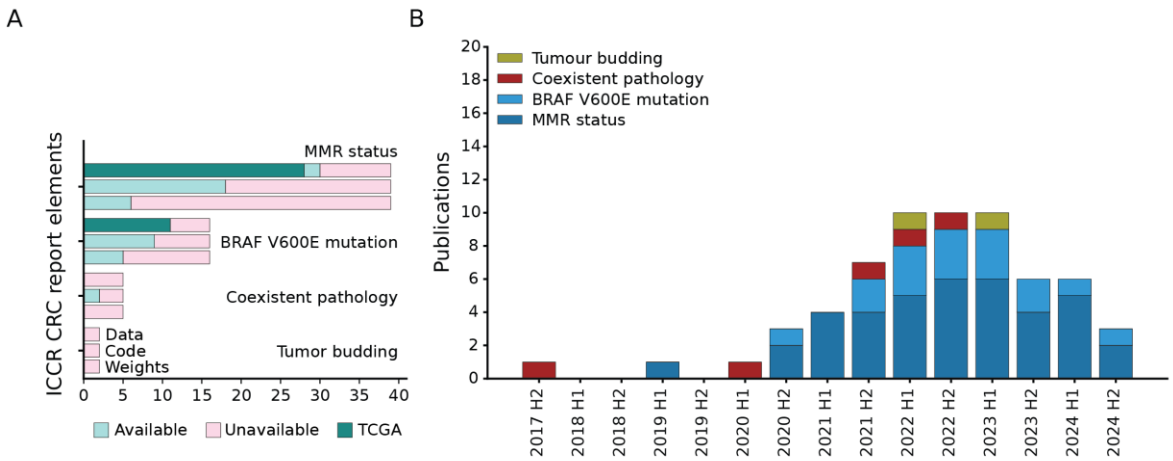

Figure 5: A) Number of articles covering a specific non-core element of the ICCR guidelines, split by the data, code, and weight availability for each topic. B) Number of published articles over time, aggregated by half-year. Not shown are elements where no publications passed exclusion criteria.

The non-core elements of the ICCR CRC dataset assess additional features that provide valuable prognostic insights and are clinically relevant but are not frequently used in patient management or not yet validated. They include measurement beyond the muscularis propria, MLH1 promoter methylation, tumor budding, coexistent pathology, MMR status, and BRAF V600E mutation. Of these six elements, only four have been addressed in the published studies included in this review. MMR status and BRAF mutation, in particular, have received significant attention, with most MMR-related (28/39) and BRAF-related (11/16) studies using TCGA as their primary data source. For these elements, code to reproduce results is often available (21/48, 44%), and some studies provide access to trained model weights (8/48, 17%). Interestingly, since similar methods are applied for MMR- and BRAF-status prediction, 14 studies investigate both biomarkers simultaneously. In contrast, tumor budding and coexistent pathology remain comparatively unexplored and overall lack publicly available data, code, or access to weights. In the following section, we examine the current research landscape for each of the non-core elements.

MMR status

The DNA mismatch repair (MMR) system is responsible to correct DNA replication mismatches[41]. The mutation of its major genes and the resulting instability of microsatellite repeat sequences is referred to Microsatellite Instability (MSI)[42]. MSI is a biomarker in CRC, guiding clinical decisions for intermediate risk Stage II and Stage IV CRCs[43,44], and represents an indicator for lynch syndrome screening - a hereditary condition characterized by mismatch repair gene mutations. Routine diagnostic testing is typically performed using Immunohistochemistry (IHC) and both ASCO and ESMO guidelines mandate dMMR testing for all new CRC diagnoses [45]. Recent DL models have demonstrated that they can predict MSI status directly from H&E slides without the need for IHC.

Study	Year	Validation Cohorts (Patients)	Best AUC	TCGA COAD/READ (FFPE) AUC**	Actionable Results	External Validation	Data	Code	Weights
Kather[46]	2019	3 (360-378)	0.84	0.77	-	✓	✓	✓	-
Schmauch[47]	2020	1 (465)	0.81	0.81	-	-	✓	✓	-
Lee[48]	2020	1 (57)	-	-	✓	-	✓	-	-
Yamashita[49]	2021	3 (100-479)	0.93	0.84	✓	✓	✓	-	-
Ke[50]	2021	1 (747)	-	-	-	-	✓	-	-
Pai[51]	2021	2 (136-230)	0.852	-	✓	-	-	-	-
Lee[52]	2021	3 (91-500)	0.97	0.89	✓	✓	✓	-	-
Schrammen[53]	2021	2 (889-2448)	0.91	-	✓	✓	-	✓	-
Saillard[54]	2021	3 (47-555)	0.97	0.88	-	✓	✓	-	-
Bilal[55]	2021	3 (47-428)	0.98	0.86	-	✓	✓	-	-
Bustos[56]	2021	1 (1788)	0.9	-	✓	-	-	-	-
Echle[57]	2022	9 (51-2452)	0.96	0.91	✓	✓	✓	✓	✓
Saldanha[58]	2022	5 (632-2448)	0.89	-	-	✓	✓	✓	-
Laleh[59]	2022	2 (632-2448)	0.93	0.89	-	-	✓	✓	-
Fujii[60]	2022	3 (248-986)	0.92	-	-	-	-	✓	✓
Jiang[61]	2022	5 (78-441)	0.88	0.88	✓	✓	✓	-	-
Guo[62]	2022	1 (396)	0.76	0.76	-	-	-	-	-
Lou[63]	2022	3 (5-144)	0.94	-	✓	-	-	-	-
Teichmann[64]	2022	1 (840)	0.94	-	-	-	✓	-	-
Lv[65]	2022	1 (868)	0.92	0.92	-	-	✓	-	-
Rubinstein[66]	2022	1 (313)	0.71	0.71	-	-	✓	-	-
Ding[67]	2022	3 (94-306)	0.84	0.84	-	✓	✓	✓	-
Cao[68]	2023	2 (47-360)	0.96	0.91	✓	✓	✓	-	-
Guo[69]	2023	2 (423-1065)	0.91	0.91	✓	✓	-	-	-
Chang[70]	2023	2 (305-1579)	0.95	0.72	✓	✓	✓	✓	-
Niehues[71]	2023	2 (2190-2448)	0.94	-	✓	-	-	✓	✓
Yuan[72]	2023	1(491)	0.93	-	-	-	✓	-	-
Tsai[73]	2023	2 (628-927)	0.88	0.88	✓	✓	✓	✓	-
Wagner[74]	2023	17 (35-2452)	0.96	0.83	✓	✓	✓	✓	-
Cen[75]	2023	2 (425-1138)	0.92	0.92	-	-	✓	✓	-
Kim[76]	2023	1 (118)	0.92	-	-	-	✓	-	-
Saillard[77]	2023	4 (47-554)	0.96	0.93	✓	✓	✓	-	-
Zamanitajeddin[78]	2024	2 (47-428)	0.99	0.88	-	✓	✓	-	-
Roth[79]	2024	2 (110-632)	0.89	0.89	-	-	✓	✓	-
Liu[80]	2024	1 (352)	0.69*	-	-	-	-	-	-
Shen[81]	2024	1 (100)	0.85	0.85	-	-	✓	-	-
Gustav[82]	2024	4 (27-2039)	0.94	0.87	✓	✓	✓	✓	-
Vorontsov[83]	2024	2 (360-482)	0.96	-	-	-	✓	✓	✓
Wang[84]	2024	3 (75-437)	0.88	0.87	-	✓	✓	✓	✓

Table 7: Extracted studies for microsatellite instability. *Publication uses multimodal data to improve results, reported are pathology only, ** Results based on some form of training only on TCGA if available.

Data and Methods

Common datasets for evaluation are the publicly available TCGA COAD/READ cohorts (80.4%), the PAIP2020 challenge dataset (26.3%), and DACHS (21.05%). Of note here is that different publications use different subsets of TCGA COAD/READ so while results have been generated on some of the same slides metrics cannot be properly compared. At the same time, many papers only used one (33.3%) or two (35.3%) cohorts in their analysis and the remainder used more (29.4%).

There are mixed approaches to ground truth, as for MSI/dMMR testing, several different methods are clinically appropriate. 33.3\% relied on both IHC and PCR for the same or different datasets, 49.0% only rely on PCR, 11.8% only on IHC, 3.9% on other methods. None of the papers using IHC as gold standard are reporting whether these were single observer results, even though there is some disagreement in manual IHC dMMR protein analysis[85].

Most MSI prediction models rely on tile level embeddings that are aggregated to a slide-level score. Before that step, many approaches pre-detect the tumor area (43.6%), and then subsequently usually add an average (41.1%), majority (23.5%) or top-k (17.6%) aggregation on top. Models that did not pre-segment the tumor tissue (56.4%) commonly relied on an attention mechanism (50%) to weight tile results for the final slide-level MSI score (Supp. Table X).

Foundation models have included MSI prediction in CRC on datasets like TCGA-CRC-DX[46,86] as a benchmark task [79,83,84], which will lead to more results in this topic, though the small scope of the evaluation may not bring additional insights on clinical translatability.

<u>Clinical Employability</u>

In MSI/dMMR prediction, the DL model either needs to predict MSI as well as a diagnostic test (e.g. at least as well as IHC) or needs to be used as a screening tool. In both cases, a cutoff needs to be defined. In this review, 47.1% of papers defined some cutoff at least on a subset of the validation data, but only 13.7% demonstrated some applicable cutoff viable for screening (e.g. at least 90% sensitivity) in a clinical setting. One method is clinically approved as a pre-screening tool with CE-IVD certification[77], yet so far, no tool can reach the same specificity as IHC or PCR testing while maintaining 95% sensitivity (Saillard et al.[77] 46% Spec. at 98% Sens., Wagner et al.[74] 61% Spec. at 95% Sens.).

The current state of papers therefore indicates that the MSI category is advanced, yet tools can only be employed as screening tools with gold-standard testing necessarily following the screening as the specificity is comparatively low. Therefore, either DL models can be further improved to have similar sensitivity and specificity as gold-standard testing, or this step necessarily needs to be deferred to a non-H&E evaluation.

BRAF V600E

BRAF mutation testing plays an important role in Lynch syndrome identification after MSI testing. On the other hand, microsatellite stable BRAF mutant colorectal cancers are particularly aggressive, yet the mutation also increases responsiveness to EGFR inhibitor therapy[87]. The non-core element in the CRC ICCR guidelines specifically concerns the oncogenic V600E mutation, which represents approximately 90% of BRAF mutant cases in a large trial[88]. Notably, BRAF and MSI cases share similarities in both morphological features and epidemiological patterns[89].

Authors [citation]	Year	Validation Cohorts (Patients)	Best AUC	TCGA COADREAD (FFPE) AUC	Actionable Results	External Validation	Data	Code	Weights
Fu[90]	2020	1 (146)	0.571	0.571	-	-	✓	✓	✓
Bilal[55]	2021	1 (502)	0.79	0.79	-	-	✓	-	-
Schrammen[53]	2021	1 (2075)	0.82	-	✓	-	-	✓	-
Chen[91]	2022	1 (414)	0.729	0.729	-	-	✓	-	-
Fujii[60]	2022	3 (248-986)	0.859	-	✓	-	-	✓	✓
Guo[62]	2022	1 (396)	0.897	0.897	-	-	-	-	-
Laleh[92]	2022	2 (534-2081)	0.86	0.808	-	✓	✓	✓	-
Saldanha[93]	2022	4(500-2075)	0.77	-	-	✓	✓	✓	-
Teichmann[64]	2022	1 (268)	0.75	0.75	-	-	✓	-	-
Cen[75]	2023	2 (500-1026)	0.78	0.78	-	✓	✓	✓	-
Guo[69]	2023	2 (500-1026)	0.8	0.77	✓	✓	-	-	-
Niehues[71]	2023	2 (1478-2081)	0.82	-	-	✓	-	✓	✓
Tsai[73]	2023	2 (591-906)	0.71	0.71	✓	✓	✓	✓	-
Wagner[74]	2023	9 (107-2081)	0.88	0.78	-	✓	✓	✓	-
Zamanitajeddin[78]	2023	1 (502)	0.79	0.79	-	-	✓	-	-
Wang[84]	2024	2 (233-559)	0.732	0.732	-	✓	✓	✓	✓

Table 8: Extracted studies for BRAF V600E mutation.

<u>Data and Methods</u>

As there is considerable overlap between the setup, task, and patients with MSI and BRAF mutation, publications frequently predict both. Therefore, different subsets of the

TCGA cohort are the also most prevalent datasets (75%), followed by the DACHS cohort (31.3%).

Most approaches apply stain normalization (68.8%) and tile extraction (100%) as preprocessing steps. Many workflows (31.3%) subsequently identify tumor tissue using a tissue classification algorithm[55,64,69,75,78]. The majority of publications implement MIL-based approaches (81.3%), differing in their tile aggregation approaches (averaging: 23.1%, top k tiles: 30.7%, other: 30.9%). These MIL-based approaches are commonly built on either CNN-based (31.3%) or attention-based (31.3%) architectures. Recent publications are showcasing foundation models (18.8%) with either averaging-based [90,91] or attention-based[84] aggregation strategies.

<u>Clinical Employability</u>

Only 43.8% of publications conducted an external validation of their algorithm, while 56.3% utilized more than one cohort. Openly available datasets were included in 68.8% of all studies. The code is available in 62% of all publications, and model weights are only available in 31.25% of studies. Furthermore, most studies did not differentiate between $BRAF^{V600E}$ mutations and all other BRAF mutations (87.5%). In contrast, Fuji et al. utilized a $BRAF^{V600E}$-specific cohort, whereas Tsai et al. reported both $BRAF^{V600E}$-specific and unspecific results.

Only four studies reported actionable results[53,60,69,73], meaning that a label ($BRAF^{WT}$/$BRAF^{mut}$) rather than a mutation probability was assigned to each WSI. Only two studies report prediction sensitivities higher than 90% [53]. Of those, only one study reached an accuracy higher than 0.5 [53].

Of all 16 included publications, only Guo et al.[69] and Schrammen et al.[25] report fully integrated pipelines with results based on a clinically appropriate cutoff (sensitivity >90%). However, Guo et al.[69] neither shared code nor weights, and Schrammen et al.[25] did not use any openly available datasets on which results could be reproduced.

The current state of $BRAF^{V600E}$ prediction is unsolved and mostly serves as a validation task for multi-purpose models. Almost all (88%) studies reporting $BRAF^{mut}$-status report MSI-status as well. As those features are highly correlated, it is expected that models will

perform similarly for both tasks. However, there exist cases that are non-correlating, which are especially interesting to pathologists[94]. None of the studies investigated performance in such cases. Furthermore, most studies do not utilize a clinically appropriate cutoff, making it hard to judge their performance in a clinical setting.

Coexistent Pathology

Under coexistent pathology, all other identified lesions should be listed. This could be additional synchronous carcinomas, polyps or e.g. IBD, other dysplastic normal tissue, subsumed under other lesions. Published work usually focuses on precursor lesions or the correct classification of polyps, but those models could be applied to e.g. correctly classify additional polyps in the same case. Additionally, because of the diversity of possible findings within this section, the models that have been developed so far only focus on one of the items that should be reported.

Study	Year	Task	Classes	Validation Cohorts (WSI)	Best Results	External Validation	Data	Code	Weights
Korbar[95]	2017	Polyp class.	NORM, HP, SSP, SA, TA, TVA	1 (239)	0.96, 0.81, 0.82, 0.90, 0.90, 0.92 Sens. 0.90, 0.90, 0.95, 0.93, 0.96, 0.95 ACC	✓	-	-	-
Wei[96]	2020	Polyp class.	TA, TVA, HP, SSA, (NORM)	2 (157-238)	0.89, 0.97, 0.82, 0.78 Sens. 0.95, 0.95, 0.96, 0.98 Spec.	✓	-	✓	-
Perlo[97]	2021	Dysplasia grading	HP, LG, HG	1 (72)	0.80, 0.73, 0.78 Sens. 0.99, 0.78, 0.88 Spec. 0.90, 0.76, 0.83 ACC	-	✓	-	-
Neto[98]	2022	Dysplasia grading	NORM, LG, HG	3 (100-1009)	1.0 Sens. 0.99 ACC	✓	✓	-	-
Soldatov[27]	2022	Polyp class.	NORM, TA, VA	1 (357)	0.55, 0.32, 0.28 Sens. 0.98, 0.79, 0.96 Spec. 0.84, 0.66, 0.87 ACC	-	-	-	-

Table 9: Extracted studies for response to coexistant pathology. NORM, Normal tissue; HP, Hyperplastic Polyp; TA, Tubular Adenoma; TVA, Tubulo-Villous Adenoma; SSP, Sessile Serrated Polyp; SA, Serrated Adenoma; SSA, Sessile Serrated Adenoma; LG, low grade dysplasia; HG, high grade dysplasia,

Data and Methods

The reviewed studies utilized datasets derived from biopsy or polyp resections rather than oncologic resection specimens. Across studies, there is lack of access to the development datasets.

It is important to highlight that the same task has different GT across studies. While only two studies address dysplasia grading, they incorporate different classes. Perlo et al.[97] add Hyperplastic Polyp (HP) to their dysplasia categories, which is not used by Neto et al.[98]. Similar observations apply for polyp classification.

Slide-level classification is achieved by Multiple Instance Learning by Neto et al.[98], while Perlo et al.[97] aggregated tile classification results from a CNN. Wei et al.[96] defined class-specific thresholds to categorize whole slide images (WSIs).

<u>Clinical Employability</u>

Variability in ground truth definitions limits model comparability and clinical integration. Differences in annotation standards hinder consistent evaluation, while a lack of publicly available datasets and independent validation cohorts restricts reproducibility. Only Neto et al.[98] provide accessible validation data, but broader external validation remains scarce. Existing studies focus on polypectomy specimens rather than oncologic resection samples, since the studies are designed for application to polyp classification in the colorectal cancer screening workflow. While existing models can assess isolated polyps, their performance on primary tissue slides, which encompass all layers of the colon, and thereby unseen tissue types, is uncertain.

Furthermore, several coexisting pathologies remain unexplored. For instance, no studies were identified that automatically evaluate (IBD) using H&E-stained slides. While Rymarczyk et al.[99] present a model to score Crohn's disease and ulcerative colitis, it relies on prior classification into one of these two categories, highlighting a gap to achieve automatic reporting of this section of the guidelines.

Tumor Budding

Tumor budding (TB) is defined as presence of small clusters of up to four cancer cells at the tumor invasive front in a tissue section (Lugli et al., 2020). TB assessment in cancer follows the standardized guidelines of the International Tumor Budding Consensus

522 Conference (ITBCC) and is performed in hotspot regions (0.785 mm^2) of H&E-stained
523 tissue sections[100]. This scoring has clinical implications, particularly for pT1 and stage II
524 CRC patients, as high TB is associated with poor prognosis [101,102]. Identifying and counting
525 TBs on H&E slides is time-consuming and prone to inter-observer variability making it an
526 ideal task to be automated by DL models.

Study	Year	Validation cohorts (WSI)	Best Results	External validation	Tumor bud detection and count	Data	Code	Weights
Lu[103]	2022	1 (40)	0.96 AUC 0.94 Sens. 0.83 Spec. 0.86 PPV	-	✓	-	-	-
Bokhorst[104]	2023	2 (15-18)	0.98 Sens. 0.44 PPV 0.58 F1	✓	✓	-	-	-

527 **Table 10:** Extracted studies for tumor budding.

528 Data and Methods

529 Most published studies on AI-assisted TB assessment rely on IHC staining, patch-level,
530 single-center training and validation of the model. After exclusion criteria, two studies
531 following the standardized ITBCC assessment pipeline were left (Table X).

532 The study by Lu, et al.[103] uses faster RCNN [105] for bud detection. Ground truth was
533 established manually on H&E tissue slides and the model's evaluation time was
534 significantly faster than manual assessment.
535 On the other hand, Bokhorst et al.[104] created a pipeline integrating a U-Net[106] model for
536 tissue segmentation and HoVer-Net[107] for nuclei detection at the invasive front. The
537 tumor bulk and invasive front were identified using a convex-hull algorithm. Their strategy
538 was validated on data from four different medical centers.
539 Clinical Employability

540 Both models present different approaches in TB detection and counting in H&E WSIs that
541 seem promising to reduce time and inter-observer variability among pathologists.
542 However, there are a number of limitations to be addressed before being able to deploy
543 such models in the diagnostic routine, such as small validation datasets, an

overestimation of TB count especially in necrotic areas, and an inability to distinguish pseudo budding [108].

4. Discussion

In this large systematic review, we have analyzed published work on deep- and machine learning for digital pathology image analysis by putting each study into context of the ICCR CRC guidelines. Particular focus was put on whether studies could be implemented in a clinical workflow to assist pathologists by automating the CRC pathology report. This work has shown substantial gaps and numerous elements of the ICCR CRC guidelines which have not been addressed by any study. Moreover, studies often do not publish code or weights such that they could be reproduced or implemented in an experimental workflow, and data is rarely shared as well.

One of the main observations from our analysis is that despite the critical role of core elements of the ICCR CRC guidelines for clinical decision making, most research efforts focus on automating the reporting of non-core elements. This discrepancy highlights the challenge in aligning the research objectives of the CPath community with the practical needs of the clinic. Many core tasks are not simple, and some are not investigated at all. Instead, they are often overlooked due to a perceived lack of methodological novelty, despite being central to real-world diagnostic workflows.

In this context, it is important to acknowledge that the recent advances in AI have powered the development of chatbots[109], or computer vision tools that can work as an end-to-end diagnostic system or even predict patient survival[110]. While these models offer the appeal of simplified outputs by bypassing intermediate diagnostic variables, they risk overlooking the interpretability and clinical validation that the elements of the ICCR guidelines offer. In contrast, building trustworthy and reliable models for ICCR-defined elements aligns more closely with current diagnostic practice. Such an approach may facilitate adoption by pathologists and laboratories, as more transparent model outputs can be verified and causally understood[111]. On the other hand, in established diagnostic criteria, it can take decades to alter established criteria or add additional ones, such that end-to-end prognostic models may not have a chance to ever be adopted as standard-of-care.

Another pressing issue is the generalized lack of multi-cohort evaluation. Models are often developed and validated on single-institution datasets or within closed consortia, which limits generalizability. Differences in staining protocols, scanning equipment, and population demographics are known to significantly affect performance, and are a source of bias[112,113], yet multi-site evaluation that would better reflect real-world conditions is rarely done, mostly due to data access barriers. As a result, many studies use TCGA data, but H&E slides in TCGA were never selected for diagnostic evaluation. Therefore, validation results on TCGA have only limited transferability to other real-world cohorts. This review excluded studies which could not report slide level results, but frequently research groups do not have access to clinical data or whole slide annotations to verify their results. Efforts such as BIGPICTURE [114] could help facilitate more clinically relevant research by making such data available.

Challenges have been crucial in pushing the boundaries of deep learning research in computational pathology[36,115,116]. Even after challenges have been completed, a number of studies still use the provided benchmark for further evaluation of new approaches. Yet, these challenges and benchmarks focus on a specific problem which may not be translatable to a diagnostic element and often do not require methods to develop deployable tools but rather highly specific methods to beat the benchmark. Moreover, inference time, economic considerations and required infrastructure are usually not addressed either, such that these challenges often are only useful for deep learning researchers, but not for clinicians and laboratory staff. On a positive note, challenges provide a standardized evaluation dataset such that methods can be compared, whereas even validation on TCGA is often done non-transparently without publishing the used data split, list of included slides in train and test set, or the exclusion criteria.

This review focused on the microscopic evaluation of CRC resection specimen and exclusively on H&E based evaluation, yet a diagnostic report would also contain macroscopic information or use other data modalities. Macroscopic analysis includes taking photos of the specimen and textual information exists, so theoretically, deep learning models could be developed for this purpose, too. Deep learning-based analysis of IHC, while more expensive since additional stainings are necessary, may be more reproducible and would still be automatically assessed[117]. Novel machines analyzing

605 tissue before embedding could also further pave the way for additional deep learning
606 tools improving diagnostic routine by creating different imaging data [118].

607 Regardless, the major results of this review are the identification of several gaps in the
608 ICCR criteria that current research has not yet covered. Table X gives an overview over all
609 topics not yet comprehensively addressed by current work. For each topic, a potential
610 approach is outlined, together with the required data and the potential benefit from
611 automatically analyzing this element.

Core Elements			
What	**How**	**Data**	**Benefits**
Histological Subtype	Classifying histological subtypes could be feasible using slide-level aggregation models to capture relevant morphological patterns. Although some studies have predicted individual subtypes, no existing algorithm differentiates between all subtypes. Rare subtypes could benefit from anomaly detection as a screening tool.	Evaluation data needs to sufficiently cover all subtypes, which may be challenging due to rarity of some subtypes (ICCR Citation). Given a heterogeneous dataset, a pretrained foundation model could reduce data requirements.	Many subtypes require a percentage (e.g. at least 50% mucinous) of the tumor volume for subtype assigment (WHO 2019). Computer aided quantification improves reproducibility and can highlight rare morphologies or calculate ratios.
Lymphatic and venous invasion (LVI)	Detecting LVI on H&E can be challenging, and pathologists often rely on elastin stains to improve accuracy (Kirsch 2013). Simple MIL models may fall short, as contextual cues, such as an artery without a nearby vein, aid detection (WHO 2019). Leveraging the spatial relationship of vessels and tumor regions may be more promising.	ICCR CRC guidelines differentiate between intramural and extramural venous invasion. Therefore, datasets should include diverse vessel sizes, invasion patterns, and anatomical locations.	LVI is a critical risk factor for stage I/II colorectal cancer (Rönnow 2022, skancke 2019), but suffers from low interobserver agreement, particularly on H&E (Kirsch 2013). Sensitive DL models could improve diagnosis in early-stages.
Margin Status	Margin status is typically assessed only when tumor is macroscopically identified within 30 mm of the resection margins. In such cases, slides from the margin are examined for tumor presence. A simple tile-based classifier may be sufficient for this task, although identifying the cutting structure could be	Generally, in most pathology archives, many blocks covering resection margins should be available, but a positive resection margin is rare. Actively annotating the cutting edge may be helpful for correct identification.	This procedure is often highly relevant for interoperative consultation when pathologists assess cryosections to assess whether a tumor has been completely removed.

	necessary to estimate the tumor-to-margin distance.		
Tumor Deposits	Tumor deposits are located in the fatty tissue on primary or lymph node WSI. To distinguish them from lymph node metastases, the boundary complexity could help, and primary tissue should not be misclassified either.	A diverse dataset including tumor deposits, lymph node metastases, and samples with and without primary tissue on the same section.	Detecting tumor deposits within fatty tissue is part of pN-staging, and leads to pN1 classification even without lymph node metastases.
Distant metastasis (pM)	The metastasis needs to be correctly identified regardless of the tissue it is in. A tile-based classifier could already be sufficient to detect the metastasis and confirm whether the morphology indeed matches a colorectal cancer morphology.	A dataset spanning tissue from multiple organs including healthy examples and samples with other pathologies that are not CRC metastases. This would help the model learn to identify metastasis and not other changes in healthy tissue.	pM is often only a confirmatory step, as the tissue with the believed metastasis has then already been extracted. But for a end-to-end second opinion by an AI model, pM is a central block of CRC diagnosis.
Neuroendocrine confirmation	For neuroendocrine neoplasms, the CRC ICCR guidelines ask for specific markers (synaptophysin and chromogranin A), and a Ki-67 proliferation index, which could also be predicted individually, though detecting the subtypes successfully would be a first step.	As neuroendocrine neoplasms are quite rare, even among colorectal neoplasms (Kooyker 2019), it may be difficult to gather enough examples to train a classifier and have the ground truth for both the stains and the Ki-67 index.	Neuroendocrine neoplasms are rare but have poor prognosis (Shafqat 2015). However, identifying neuroendocrine neoplasms in the first place is unsolved.
Non-Core Elements			
What	**How**	**Data**	**Benefits**
Measurement of invasion beyond muscularis propria	To measure the invasion depth cases with tumor invasion beyond the muscularis propria (MP) must first be identified. Then, a tissue segmentation model can detect tumor regions beyond the MP and measure the orthogonal distance to it.	The necessary samples for this type of project are common, as any CRC > pT2 can be included. Pathologist annotations for the continuous measurement would be necessary on a sufficiently large cohort.	Evidence suggests (Merkel 2001, Bori 2009) that pT3 can be stratified into pT3a and pT3b with additional prognostic value. A large-scale investigation using an automatic method could provide further insight.
MLH1 promoter methylation	Given the current state of MSI prediction where high specificity with clinical grade sensitivity cannot be reached, trying to make an even more specific prediction than the general MSI label, is a high-risk project.		

Table 11: Addressing the remaining ICCR CRC guideline elements.

This overview could be a starting point for a number of investigations, yet it is important to mention that each element may have its own set of specific problems that only arise once researchers investigate the element.

5. Limitations

Limiting the scope of the review to papers with slide-level results and automated analysis ignores a sizable number of papers that could likely be translated to working WSI level approaches with little additional work. Yet, these excluded articles are also not less important for the progress in the field. Moreover, the employed search terms potentially excluded relevant papers if they did not include the indicated terms in the abstract. By selecting ICCR CRC guidelines over alternatives like CAP, we emphasized certain topics while omitting emerging biomarkers such as tumor-stroma ratio, tumor border configuration, and immune response. Additionally, the focus on H&E-based algorithms ignores immunohistochemistry-based approaches, which represent promising alternatives in CRC diagnostics.

6. Strengths

This systematic review includes publications on deep- and machine-learning AI-models for image analysis in pathology according to elements of the ICCR CRC dataset[7] - a globally recognized, evidence-based and standardized pathology reporting guideline. It provides the readers with the current landscape of research in CPath towards automating routine tasks of pathologists during CRC assessment according to ICCR guideline. In the results section we provide a general analysis on the trend of publications over the years and per each element of the guideline. We also present information on the data, code and weight availability among included publications.

Following the general overview, we dissect publications individually, summarize their key results in separate tables as well as critically evaluate the clinical readiness of the proposed models, for each core and non-core element of the ICCR CRC guideline.

Moreover, to guide future research efforts, we present a dedicated table containing all underexplored core and non-core elements, where we propose a training approach, the dataset types and the anticipated benefits of utilizing such models.

Finally, we emphasize practical aspects for researchers, such as the necessity for transparency in computational requirements, availability of code, weights and data - essential for reproducibility and improvement of AI tools with clinical applicability.

All in all, this review should facilitate alignment between future research efforts in CPath and clinical practices as well as encourage new studies to fill current gaps and underexplored research areas.

7. Conclusion

This systematic review provides a comprehensive overview of the current state of CPath research to address elements of the diagnostic report in CRC. It reveals the limited maturity of the field and lack of alignment between research and clinical practices and emphasizes the need for developing approaches for various gaps while evaluating existing methods in clinical contexts. The findings could also lead to new benchmarks for novel models that assess entire diagnostic workflows rather than isolated tasks. Finally, this review underscores the remarkable complexity of CRC diagnosis performed by pathologists, which represents just one of many diseases they routinely evaluate every day.

8. Acknowledgements

This work was made possible via funding from several sources. E.B. and A.L.F. were funded by the Swiss National Science Foundation (CRSII5_193832). A.K. was funded by the Center for Artificial Intelligence in Medicine (CAIM), University of Bern. M.G. and J.F.C. was funded by the Swiss cancer league (KFS-5786-02-2023-R). J.H. was funded by the Institute of Tissue Medicine and Pathology, University of Bern. R.M. was funded by Swiss National Science Foundation (31003A_166578/1) and the Swiss Government Excellence Scholarship (ESKAS, nr. 2021.0019 / Kosovo / OP). J.G.B. was funded by the Swiss National Science Foundation (10.000.619).

We would also like to thank Tanya Karrer, University Library of Bern, for her valuable help in the conceptual development of this review.

9. Competing interests

Inti Zlobec acts as scientific advisor for Aiforia.

1. Hanna, M. G. *et al.* Integrating digital pathology into clinical practice. *Modern Pathology 2021 35:2* **35**, 152–164 (2021).
2. Eloy, C. *et al.* Artificial intelligence–assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies. *Virchows Archiv* **482**, 595–604 (2023).
3. Frei, A. L. *et al.* Pathologist Computer-Aided Diagnostic Scoring of Tumor Cell Fraction: A Swiss National Study. *Modern Pathology* **36**, 100335 (2023).
4. Bertram, C. A. *et al.* Computer-assisted mitotic count using a deep learning–based algorithm improves interobserver reproducibility and accuracy. *Vet Pathol* **59**, 211–226 (2022).
5. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209–249 (2021).
6. Nagtegaal, I. D. *et al.* The 2019 WHO classification of tumours of the digestive system. *Histopathology* **76**, 182 (2019).
7. Loughrey, M. B. *et al.* Dataset for Pathology Reporting of Colorectal Cancer: Recommendations from the International Collaboration on Cancer Reporting (ICCR). *Ann Surg* **275**, 549–561 (2022).
8. Burgart, L. J., Chopp, W. V & Jain, D. Protocol for the Examination of Resection Specimens From Patients With Primary Carcinoma of the Colon and Rectum With guidance from the CAP Cancer and CAP Pathology Electronic Reporting Committees. (2022).
9. Moher, D. *et al.* Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann Intern Med* **151**, 264–269 (2009).
10. Kellermeyer, L., Harnke, B. & Knight, S. Covidence and Rayyan. *J Med Libr Assoc* **106**, 580 (2018).
11. Song, A. H. *et al.* Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering 2023 1:12* **1**, 930–949 (2023).

12. Brierley, J., Gospodarowicz, M. D. & Wittekind, C. T. TNM Classification of Malignant Tumors International Union Against Cancer. 8th. Oxford, England: Wiley; 2017. *Wiley* pp57-62 (2017).

13. Ong, M. L. H. & Schofield, J. B. Assessment of lymph node involvement in colorectal cancer. *World J Gastrointest Surg* **8**, 179 (2016).

14. Chuang, W.-Y. *et al.* Identification of nodal micrometastasis in colorectal cancer using deep learning on annotation-free whole-slide images. *Modern Pathology* **34**, 1901–1911 (2021).

15. Mainovskaya, O. A. *et al.* Development and application of artificial intelligence for the detection of the lymph nodes metastases in colorectal cancer. *Koloproktologia* **21**, 49–59 (2022).

16. Tsuneki, M. & Kanavati, F. Deep learning models for poorly differentiated colorectal adenocarcinoma classification in whole slide images using transfer learning. *Diagnostics* **11**, (2021).

17. Bándi, P. *et al.* Continual learning strategies for cancer-independent detection of lymph node metastases. *Med Image Anal* **85**, (2023).

18. Khan, A. *et al.* Computer-Assisted Diagnosis of Lymph Node Metastases in Colorectal Cancers Using Transfer Learning With an Ensemble Model. *Modern Pathology* **36**, (2023).

19. Tan, L. *et al.* Colorectal cancer lymph node metastasis prediction with weakly supervised transformer-based multi-instance learning. *Med Biol Eng Comput* **61**, 1565–1580 (2023).

20. Kindler, C., Elfwing, S., Öhrvik, J. & Nikberg, M. A Deep Neural Network–Based Decision Support Tool for the Detection of Lymph Node Metastases in Colorectal Cancer Specimens. *Modern Pathology* **36**, (2023).

21. Giammanco, A. *et al.* Fast-Track Development and Multi-institutional Clinical Validation of an Artificial Intelligence Algorithm for Detection of Lymph Node Metastasis in Colorectal Cancer. *Modern Pathology* **37**, (2024).

22. Huang, Z. *et al.* A pathologist–AI collaboration framework for enhancing diagnostic accuracies and efficiencies. *Nat Biomed Eng* (2024) doi:10.1038/s41551-024-01223-5.

23. Chandler, I. & Houlston, R. S. Interobserver agreement in grading of colorectal cancers-findings from a nationwide web-based survey of histopathologists. *Histopathology* **52**, 494–499 (2008).

24. Rathore, S. *et al.* Segmentation and grade prediction of colon cancer digital pathology images across multiple institutions. *Cancers (Basel)* **11**, (2019).

25. Schrammen, P. L. *et al.* Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J Pathol* **256**, 50–60 (2022).

26. Chen, S. *et al.* Automatic Tumor Grading on Colorectal Cancer Whole-Slide Images: Semi-Quantitative Gland Formation Percentage and New Indicator Exploration. *Front Oncol* **12**, (2022).

27. Soldatov, S. A. *et al.* Deep Learning Classification of Colorectal Lesions Based on Whole Slide Images. *Algorithms* **15**, (2022).

28. Ueno, H. *et al.* New criteria for histologic grading of colorectal cancer. *American Journal of Surgical Pathology* **36**, 193–201 (2012).

29. Levy, J., Haudenschild, C., Barwick, C., Christensen, B. & Vaickus, L. Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks. *Pac Symp Biocomput* **26**, 285–296 (2021).

30. Kumar, N. *et al.* Computer-extracted features of nuclear morphology in hematoxylin and eosin images distinguish stage II and IV colon tumors. *J Pathol* **257**, 17–28 (2022).

31. Pei, Z., Zhang, D. & Shao, W. Efficient Metric Learning with Graph Transformer for Accurate Colorectal Cancer Staging. in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* 1–6 (2022). doi:10.1109/BHI56158.2022.9926858.

32. Liebig, C. *et al.* Perineural invasion is an independent predictor of outcome in colorectal cancer. *Journal of Clinical Oncology* **27**, 5131–5137 (2009).

33. Huh, J. W., Kim, H. R. & Kim, Y. J. Prognostic value of perineural invasion in patients with stage II colorectal cancer. *Ann Surg Oncol* **17**, 2066–2072 (2010).

34. Han, C. H. & Kwak, J. T. A hybrid computational pathology method for the detection of perineural invasion junctions. *https://doi.org/10.1117/12.2610756* **12039**, 215–219 (2022).

35. Jung, J., Kim, E., Lee, H., Lee, S. H. & Ahn, S. Automated Hybrid Model for Detecting Perineural Invasion in the Histology of Colorectal Cancer. *APPLIED SCIENCES-BASEL* **12**, (2022).

36. Choi, J., Lee, K., Jeong, W.-K. & Chun, S. Y. PAIP2021: Perineural Invasion in Multiple Organ Cancer (Colon, Prostate, and Pancreatobiliary tract). in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* (2021). doi:10.5281/ZENODO.4575424.

37. Ju, Y. H. *et al.* Artificial intelligence recognition of pathological T stage and tumor invasion in rectal cancer based on large panoramic pathological sections. *INTELLIGENT MEDICINE* **2**, 141–151 (2022).

38. Langer, R. & Becker, K. Tumor regression grading of gastrointestinal cancers after neoadjuvant therapy. *Virchows Archiv* **472**, 175–186 (2018).

39. Zhang, D. *et al.* Using Multi-Scale Convolutional Neural Network Based on Multi-Instance Learning to Predict the Efficacy of Neoadjuvant Chemoradiotherapy for Rectal Cancer. *IEEE J Transl Eng Health Med* **10**, (2022).

40. Amin, M. B. *et al. AJCC Cancer Staging Manual*. vol. 1024 (Springer, 2017).

41. Kunkel, T. A. & Erie, D. A. DNA mismatch repair. *Annu Rev Biochem* **74**, 681–710 (2005).

42. Boland, C. R. & Goel, A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology* **138**, 2073-2087.e3 (2010).

43. Argilés, G. *et al.* Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up y on behalf of the ESMO Guidelines Committee. *Annals of Oncology* **31**, 1291–1305 (2020).

44. Cervantes, A. *et al.* Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up 5 behalf of the ESMO Guidelines Committee. *Annals of Oncology* **34**, 10–32 (2023).

45. Stoffel, E. M. *et al.* Hereditary Colorectal Cancer Syndromes: American Society of Clinical Oncology Clinical Practice Guideline Endorsement of the Familial Risk–Colorectal Cancer: European Society for Medical Oncology Clinical Practice Guidelines. *Journal of Clinical Oncology* **33**, 209 (2014).

46. Kather, J. N. *et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine 2019 25:7* **25**, 1054–1056 (2019).

47. Schmauch, B. *et al.* A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications 2020 11:1* **11**, 1–15 (2020).

48. Lee, H. *et al.* Two-Stage Classification Method for MSI Status Prediction Based on Deep Learning Approach. *APPLIED SCIENCES-BASEL* **11**, (2021).

49. Yamashita, R. *et al.* Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. *Lancet Oncol* **22**, 132–141 (2021).

50. Ke, J. *et al.* Identifying patch-level MSI from histological images of Colorectal Cancer by a Knowledge Distillation Model. in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1043–1046 (2020). doi:10.1109/BIBM49941.2020.9313141.

51. Pai, R. K. *et al.* Development and initial validation of a deep learning algorithm to quantify histological features in colorectal carcinoma including tumour budding/poorly differentiated clusters. *Histopathology* **79**, 391–405 (2021).

52. Lee, S. H., Song, I. H. & Jang, H.-J. Feasibility of deep learning-based fully automated classification of microsatellite instability in tissue slides of colorectal cancer. *Int J Cancer* **149**, 728–740 (2021).

53. Schrammen, P. L. *et al.* Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *Journal of Pathology* **256**, 50–60 (2022).

54. Saillard, C. *et al.* Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. *Nat Commun* **14**, (2023).

55. Bilal, M. *et al.* Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health* **3**, e763–e772 (2021).

56. Bustos, A. *et al.* XDEEP-MSI: Explainable bias-rejecting microsatellite instability deep learning system in colorectal cancer. *Biomolecules* **11**, (2021).

57. Echle, A. *et al.* Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application. *ESMO Open* **7**, (2022).

58. Saldanha, O. L. *et al.* Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat Med* **28**, 1232–1239 (2022).

59. Ghaffari Laleh, N. *et al.* Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med Image Anal* **79**, (2022).

60. Fujii, S. *et al.* Rapid Screening Using Pathomorphologic Interpretation to Detect BRAFV600E Mutation and Microsatellite Instability in Colorectal Cancer. *Clinical Cancer Research* **28**, 2623–2632 (2022).

61. Jiang, W. *et al.* Clinical actionability of triaging DNA mismatch repair deficient colorectal cancer from biopsy samples using deep learning. *EBioMedicine* **81**, (2022).

837 62. Guo, Y. *et al.* Learn to Estimate Genetic Mutation and Microsatellite Instability
838 with Histopathology H&E Slides in Colon Carcinoma. *Cancers (Basel)* **14**, (2022).

839 63. Lou, J. *et al.* PPsNet: An improved deep learning model for microsatellite
840 instability high prediction in colorectal cancer from whole slide images. *Comput*
841 *Methods Programs Biomed* **225**, (2022).

842 64. Teichmann, M., Aichert, A., Bohnenberger, H., Ströbel, P. & Heimann, T. End-to-
843 end learning for image-based detection of molecular alterations in digital
844 pathology. in *International Conference on Medical Image Computing and*
845 *Computer-Assisted Intervention* 88–98 (2022).

846 65. Lv, Z., Yan, R., Lin, Y., Wang, Y. & Zhang, F. Joint region-attention and multi-scale
847 transformer for microsatellite instability detection from whole slide images in
848 gastrointestinal cancer. in *International Conference on Medical Image*
849 *Computing and Computer-Assisted Intervention* 293–302 (2022).

850 66. Rubinstein, J. C. *et al.* Deep learning image analysis quantifies tumor
851 heterogeneity and identifies microsatellite instability in colon cancer. *J Surg*
852 *Oncol* **127**, 426–433 (2023).

853 67. Ding, K., Zhou, M., Wang, H., Zhang, S. & Metaxas, D. N. Spatially aware graph
854 neural networks and cross-level molecular profile prediction in colon cancer
855 histopathology: a retrospective multi-cohort study. *Lancet Digit Health* **4**, e787–
856 e795 (2022).

857 68. Cao, R., Gu, Q., Tan, D., Wei, P. & Zheng, C. Prediction of microsatellite instability
858 of colorectal cancer using multi-scale pathological images based on deep
859 learning. in *2022 IEEE International Conference on Bioinformatics and*
860 *Biomedicine (BIBM)* 1461–1466 (2022). doi:10.1109/BIBM55620.2022.9995576.

861 69. Guo, B. *et al.* Predicting microsatellite instability and key biomarkers in colorectal
862 cancer from H&E-stained images: achieving state-of-the-art predictive
863 performance with fewer data using Swin Transformer. *Journal of Pathology:*
864 *Clinical Research* **9**, 223–235 (2023).

865 70. Chang, X. *et al.* Predicting colorectal cancer microsatellite instability with a self-
866 attention-enabled convolutional neural network. *Cell Rep Med* **4**, (2023).

867 71. Niehues, J. M. *et al.* Generalizable biomarker prediction from cancer pathology
868 slides with self-supervised deep learning: A retrospective multi-centric study.
869 *Cell Rep Med* **4**, (2023).

870 72. Yuan, X., Ruan, J. & Yue, J. Classification of Microsatellite Instability Status in
871 Slide-level Annotated Colorectal Tumors by Weakly Supervised Deep Learning. in

2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC) vol. 6 1306–1310 (2023).

73. Tsai, P.-C. *et al.* Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nat Commun* **14**, (2023).

74. Wagner, S. J. *et al.* Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. *Cancer Cell* **41**, 1650-1661.e4 (2023).

75. Cen, M. *et al.* A Novel and Efficient Digital Pathology Classifier for Predicting Cancer Biomarkers Using Sequencer Architecture. *American Journal of Pathology* **193**, 2122–2132 (2023).

76. Kim, K. *et al.* PAIP 2020: Microsatellite instability prediction in colorectal cancer. *Med Image Anal* **89**, 102886 (2023).

77. Saillard, C. *et al.* Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides. *Nature Communications 2023 14:1* **14**, 1–11 (2023).

78. Zamanitajeddin, N., Jahanifar, M., Bilal, M., Eastwood, M. & Rajpoot, N. Social network analysis of cell networks improves deep learning for prediction of molecular pathways and key mutations in colorectal cancer. *Med Image Anal* **93**, 103071 (2024).

79. Roth, B. *et al.* Low-Resource Finetuning of Foundation Models Beats State-of-the-Art in Histopathology. in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)* 1–5 (2024). doi:10.1109/ISBI56570.2024.10635695.

80. Liu, Q. *et al.* M 2 Fusion: Bayesian-Based Multimodal Multi-level Fusion on Colorectal Cancer Microsatellite Instability Prediction. in *International Conference on Medical Image Computing and Computer-Assisted Intervention* 125–134 (2023).

81. Shen, Y., Zhou, B., Xiong, X., Gao, R. & Wang, Y. G. How GNNs Facilitate CNNs in Mining Geometric Information from Large-Scale Medical Images. in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2227–2230 (2023). doi:10.1109/BIBM58861.2023.10385379.

82. Gustav, M. *et al.* Deep learning for dual detection of microsatellite instability and POLE mutations in colorectal cancer histopathology. *NPJ Precis Oncol* **8**, (2024).

83. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine 2024 30:10* **30**, 2924–2935 (2024).

84. Wang, X. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 21 (2024).

85. Klarskov, L. *et al.* Interobserver variability in the evaluation of mismatch repair protein immunostaining. *Hum Pathol* **41**, 1387–1396 (2010).

86. Kather, J. N. Histological image tiles for TCGA-CRC-DX, color-normalized, sorted by MSI status, train/test split. Preprint at https://doi.org/10.5281/ZENODO.3832231 (2020).

87. Sharma, S. G. & Gulley, M. L. BRAF Mutation Testing in Colorectal Cancer. *Arch Pathol Lab Med* **134**, 1225–1228 (2010).

88. Kopetz, S. *et al.* Molecular profiling of BRAF-V600E-mutant metastatic colorectal cancer in the phase 3 BEACON CRC trial. *Nature Medicine 2024 30:11* **30**, 3261–3271 (2024).

89. Nicolò Fanelli, G. *et al.* The heterogeneous clinical and pathological landscapes of metastatic Braf-mutated colorectal cancer. doi:10.1186/s12935-020-1117-2.

90. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* **1**, 800–810 (2020).

91. Chen, Z. H., Li, X. Y., Yang, M. M., Zhang, H. & Xu, X. S. Optimization of deep learning models for the prediction of gene mutations using unsupervised clustering. *JOURNAL OF PATHOLOGY CLINICAL RESEARCH* **9**, 3–17 (2023).

92. Ghaffari Laleh, N. *et al.* Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Med Image Anal* **79**, (2022).

93. Saldanha, O. L. *et al.* Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat Med* **28**, 1232–1239 (2022).

94. Tran, B. *et al.* Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer. *Cancer* **117**, 4623–4632 (2011).

95. Korbar, B. *et al.* Deep Learning for Classification of Colorectal Polyps on Whole-slide Images. *J Pathol Inform* **8**, 30 (2017).

96. Wei, J. W. *et al.* Evaluation of a Deep Neural Network for Automated Classification of Colorectal Polyps on Histopathologic Slides. *JAMA Netw Open* **3**, e203398–e203398 (2020).

97. Perlo, D., Tartaglione, E., Bertero, L., Cassoni, P. & Grangetto, M. Dysplasia Grading of Colorectal Polyps Through Convolutional Neural Network Analysis of Whole Slide Images. in 325–334 (2022). doi:10.1007/978-981-16-3880-0_34.

98. Neto, P. C. *et al.* iMIL4PATH: A Semi-Supervised Interpretable Approach for Colorectal Whole-Slide Images. *Cancers 2022, Vol. 14, Page 2489* **14**, 2489 (2022).

99. Rymarczyk, D. *et al.* Deep Learning Models Capture Histological Disease Activity in Crohn's Disease and Ulcerative Colitis with High Fidelity. *J Crohns Colitis* **18**, 604–614 (2024).

100. Lugli, A. *et al.* Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. *Modern Pathology* **30**, 1299–1311 (2017).

101. Grigore, A. D., Jolly, M. K., Jia, D., Farach-Carson, M. C. & Levine, H. Tumor Budding: The Name is EMT. Partial EMT. *Journal of Clinical Medicine 2016, Vol. 5, Page 51* **5**, 51 (2016).

102. Lugli, A., Zlobec, I., Berger, M. D., Kirsch, R. & Nagtegaal, I. D. Tumour budding in solid cancers. *Nature Reviews Clinical Oncology 2020 18:2* **18**, 101–115 (2020).

103. Lu, J. Q. *et al.* Development and application of a detection platform for colorectal cancer tumor sprouting pathological characteristics based on artificial intelligence. *INTELLIGENT MEDICINE* **2**, 82–87 (2022).

104. Bokhorst, J.-M. *et al.* Fully Automated Tumor Bud Assessment in Hematoxylin and Eosin-Stained Whole Slide Images of Colorectal Cancer. *Modern Pathology* **36**, (2023).

105. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans Pattern Anal Mach Intell* **39**, 1137–1149 (2017).

106. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9351**, 234–241 (2015).

107. Graham, S. *et al.* Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal* **58**, 101563 (2019).

108. Haddad, T. S. *et al.* Pseudobudding: ruptured glands do not represent true tumor buds. *J Pathol* **261**, 19–27 (2023).

109. Lu, M. Y. *et al.* A multimodal generative AI copilot for human pathology. *Nature* **634**, 466–473 (2024).

110. Kleppe, A. *et al.* A clinical decision support system optimising adjuvant chemotherapy for colorectal cancers by integrating deep learning and

pathological staging markers: a development and validation study. *Lancet Oncol* **23**, 1221–1232 (2022).

111. Plass, M. *et al.* Explainability and causability in digital pathology. *J Pathol Clin Res* **9**, 251–260 (2023).

112. Asilian Bidgoli, A., Rahnamayan, S., Dehkharghanian, T., Grami, A. & Tizhoosh, H. R. Bias reduction in representation of histopathology images using deep feature selection. *Scientific Reports 2022 12:1* **12**, 1–12 (2022).

113. Vaidya, A. *et al.* Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine 2024 30:4* **30**, 1174–1190 (2024).

114. Moulin, P., Grünberg, K., Barale-Thomas, E. & der Laak, J. van. IMI—Bigpicture: A Central Repository for Digital Pathology. *Toxicol Pathol* **49**, 711–713 (2021).

115. Graham, S. *et al.* CoNIC: Colon Nuclei Identification and Counting Challenge 2022. (2021).

116. Litjens, G. *et al.* 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience* **7**, (2018).

117. Lems, C. M. *et al.* Color Deconvolution for Color-Agnostic and Cross-Modality Analysis of Immunohistochemistry Whole-Slide Images with Deep Learning. *Proceedings - International Symposium on Biomedical Imaging* (2024) doi:10.1109/ISBI56570.2024.10635258.

118. Jorns, J. M. *et al.* Shortwave infrared imaging increases the number of lymph nodes and the number of positive lymph nodes in surgical pathology specimens – a 104-patient study. doi:10.1101/2024.11.24.24317845.