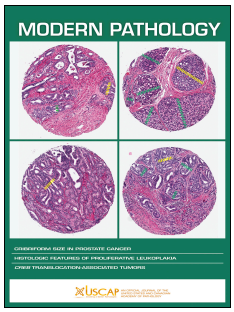
Journal Pre-proof



Fully Automated Tumor Bud Assessment in Hematoxylin and Eosin Stained Whole Slide Images of Colorectal Cancer

John-Melle Bokhorst, Francesco Ciompi, Sonay Kus Öztürk, Ayse Selcen Oguz Erdogan, Michael Vieth, Heather Dawson, Richard Kirsch, Femke Simmer, Kieran Sheahan, Alessandro Lugli, Inti Zlobec, Jeroen van der Laak, Iris D. Nagtegaal

PII: S0893-3952(23)00138-2

DOI: https://doi.org/10.1016/j.modpat.2023.100233

Reference: MODPAT 100233

To appear in: *Modern Pathology*

Received Date: 22 January 2023

Revised Date: 25 April 2023

Accepted Date: 20 May 2023

Please cite this article as: Bokhorst JM, Ciompi F, Öztürk SK, Oguz Erdogan AS, Vieth M, Dawson H, Kirsch R, Simmer F, Sheahan K, Lugli A, Zlobec I, van der Laak J, Nagtegaal ID, Fully Automated Tumor Bud Assessment in Hematoxylin and Eosin Stained Whole Slide Images of Colorectal Cancer, *Modern Pathology* (2023), doi: https://doi.org/10.1016/j.modpat.2023.100233.

Title:

Fully Automated Tumor Bud Assessment in Hematoxylin and Eosin Stained Whole Slide Images of Colorectal Cancer

Authors: John-Melle Bokhorst[a], Francesco Ciompi[a], Sonay Kus Öztürk[a], Ayse Selcen Oguz Erdogan[a], Michael Vieth[b], Heather Dawson[c], Richard Kirsch[d], Femke Simmer[a], Kieran Sheahan[e], Alessandro Lugli[b], Inti Zlobec[b], Jeroen van der Laak[a,f], Iris D. Nagtegaal[a]

[a] Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands
[b] Klinikum of pathology, Bayreuth University, Bayreuth, Germany
[c] Institute of Tissue Medicine and Pathology, University of Bern, 3008 Bern, Switzerland
[d] University of Toronto, Mount Sinai Hospital, Toronto, Canada
[e] Departments of Pathology, St Vincent's Hospital, Dublin, Ireland.
[f] Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden

Corresponding author:
John-Melle Bokhorst, +31 (024) 361 43 14, john-melle.bokhorst@radboudumc.nl,
Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands

Keywords:
Tumor budding, Colorectal cancer, automated assessment, Prognosis, Artificial intelligence, computational pathology.

Running title:

Fully Automated Tumor Bud Assessment in H&E-stained WSIs

Abstract

Tumor budding (TB), the presence of single cells or small clusters of up to four tumor cells, at the invasive front of colorectal cancer (CRC) is a proven risk factor for adverse outcomes. International definitions are necessary to reduce the interobserver variability. According to the current international guideline, hotspots at the invasive front should be counted in Hematoxylin and Eosin (H&E) stained slides. This is time-consuming and prone to interobserver variability, therefore there is a need for computer-aided diagnosis solutions.

In this paper, we report on developing an Artificial Intelligence (AI) based method for detecting tumor budding in H&E-stained whole slide images. We propose a fully automated pipeline to identify the tumor border, detect tumor buds, characterize them based on their number of tumor cells, and produce a TB density map that we use to identify the TB hot spot. The method outputs the TB count in the hotspot as a computational biomarker.

We show that the proposed automated TB detection workflow performs on par with a panel of five pathologists at detecting tumor buds, and that the hotspot-based TB count is an independent prognosticator in both the univariate and the multivariate analysis, validated on a cohort of n=981 CRC patients.

Computer-aided detection of tumor buds based on deep learning can perform on par with expert pathologists at detection and quantification of tumor buds in H&E-stained colorectal cancer histopathology slides, strongly facilitating the introduction of budding as an independent prognosticator in clinical routine and clinical trials.

Introduction

Colorectal cancer (CRC) is among the most common cancers, responsible for almost 2,000,000 new cases every year and over 900,000 deaths globally [1]. Staging of CRC is based on invasion depth (T), number of involved lymph nodes (N), and the presence of distant metastases (M) [2]. However, to accurately predict disease progression of individual patients, this staging system is insufficient and additional biomarkers are required to decide on therapeutic strategies and prevent possible under- and overtreatment. Among the most promising histological biomarkers in CRC are tumor budding (TB) [3] and poorly differentiated clusters (PDC) [4]. TB is defined as isolated single cells or small clusters of up to four tumor cells located at the invasive tumor front. In contrast, clusters of five or more tumor cells without gland formation are defined as PDCs. Both biomarkers are associated with lymph node and distant metastasis, and increased patient mortality in CRC [5].

International harmonization of TB scoring was achieved by the International Tumor Budding Consensus Conference (ITBCC) recommendations in 2016 [3]: identification of the budding hotspot (measuring 0·785 mm^2) at the invasive front of the tumor, and counting of the number of buds, results in a score that can be classified as Bd1 (0-4 buds, low budding), Bd2 (5-9 buds, intermediate budding) or Bd3 (10 or more buds, high budding). Although pan-cytokeratin immunohistochemical staining might be helpful in the recognition of TB, the definition is based on Hematoxylin and Eosin (H&E) slide scoring. Despite these clear definitions, interobserver agreement is moderate to substantial [6].

In recent years, the application of machine learning approaches based on deep learning has increased the accuracy, reproducibility, and efficiency of histopathologic slide analysis [7]. When presented with sufficient high-quality annotated training data, convolutional neural networks (CNNs), a special type of deep learning based on artificial neural networks, can learn complex histological patterns from structured data such as medical images. Current applications of artificial intelligence in computational pathology (CPATH) include recognition, segmentation, and classification of morphological structures in digital pathology whole-slide images (WSI), such as tumor (peripheral) regions or the detection and

classification of several types of cells, including tumor cells, to be used as a base for development of computational biomarkers [8].

Despite advances in the field of CPATH, computer-aided detection of tumor buds in H&E remains a challenging task, mostly due to 1) the small size of the objects to detect (i.e., the tumor buds), 2) a possible variation in phenotype due to an epithelial to mesenchymal transition [9], and 3) the very heterogeneous composition of the microenvironment in which they are located. In addition, observer variability in identifying individual buds makes collection of a large training set of TB and PDC challenging. To the best of our knowledge, to date most of the (semi-)automatic TB detection methods work on immunohistochemically stained slides (reviewed in [10]), and only a few H&E-based algorithms have been proposed [11,12], either based on single-center slides [11], limiting their generalizability, or not distinguishing TB from PDC [12].

In this study, we developed a novel fully automated pipeline for computer-aided detection and quantification of tumor buds in whole-slide images from H&E-stained CRC specimens. Following the recommendations of the ITBCC, our method automatically identifies the tumor invasive front, detects tumor buds, and quantifies them in a hot-spot driven fashion. We compared the detection performance of our computer-aided detection model with a panel of five pathologists, and we analyzed the prognostic value of our computer-based hotspot-derived TB count on a large external independent cohort or CRC cases.

Materials

In this section, we describe the datasets used in this work, namely 1) a single-center *model development* dataset, containing H&E slides with manual annotations, used to train a deep learning algorithm to segment epithelial regions; 2) a multi-centric *technical validation* dataset, used to validate the tumor-bud detection performance of the developed method; 3) a multi-centric *clinical validation* dataset, used to assess the prognostic value of the automated TB count produced by the presented method.

Model development dataset

Between one and five slides from CRC patients (n=37) with the presence of tumor budding reported during the initial sign-out were collected from the Radboud University Medical Center (Radboudumc), Nijmegen (Netherlands). In total, 60 slides were included. Only slides where the invasive front was clearly visible were selected, stained with H&E, digitized, and subsequently restained with a cytokeratin 8-18 (CK8-18) immunohistochemical (IHC) marker after scanning, according to our established protocol [13], resulting in two versions of the same slide, that were digitized using the Pannoramic P250 Flash II scanner (3D-Histech, Budapest, Hungary), at 40X magnification (spatial resolution of 0·24µm/px). On each slide, regions of interest (ROI) near the invasive front were manually selected and transferred to CK8-18 slides, where all positive cells were delineated. These annotations were transferred back to H&E slides and removed when no bud was visible in the H&E slide. Subsequently, the remaining non-epithelium pixels within the ROI were labeled as either necrosis or background. Annotations were made using the in-house developed open-source software ASAP (https://github.com/computationalpathologygroup/ASAP). The dataset was randomly split into a training (n=42) and validation set (n=18). We will refer to this dataset as Dev. An overview of the annotating procedure can be found in Figure 1A.

Technical validation dataset

For technical validation, i.e., the comparison between detection performance of the proposed CAD system and a panel of pathologists, we included a set of n=15 whole slide images selected from four

different centers: Dublin University Hospital (Dublin, Ireland), Bayreuth University Hospital (Bayreuth, Germany), Bern University, Institute of Pathology (Bern, Switzerland), Mount Sinai Hospital (Toronto, Canada). Cases were selected based on the presence of high budding reported in the original diagnostic report. Glass slides were stained with H&E at the pathology laboratory of each center, therefore generating variation in the H&E staining, scanned, re-stained with CK8-18 immunohistochemical marker and re-scanned at Radboudumc using a Pannoramic P250 Flash II scanner (3D-Histech, Budapest, Hungary), at 40X magnification. Manual annotations were made as described above. We will refer to this dataset as Val-t.

Clinical validation dataset

For clinical validation, i.e., the analysis of the prognostic value of an automated TB score, n=557 CRC patients from Radboudumc and n=568 CRC patients from Mount Sinai Hospital, Toronto (Canada) were included. Based on the pathologists' assessment of multiple sections, the clinical validation set was established by visual selection of a single slide per case in line with the ITBCC guidelines. Radboudumc slides were scanned using the same 3DHistech scanner used in Val-t; the Canadian cases were scanned with Aperio AT2 scanner (Leica Biosystems, Buffalo Grove, IL, USA). All slides were scanned at 40X magnification yielding a spatial resolution of 0·24um/px. In the Canadian cohort a tumor bud count was established by an expert, according to ITBCC recommendations; n=144 patients received neoadjuvant treatment and were therefore excluded from this study. As a result, n=981 CRC patients were included in the clinical validation, for a total of n=1125 WSIs.

We will refer to these sets in the clinical validation dataset as Cohort C and D for the Canadian and the Radboudumc cohort, respectively.

Method

Our deep learning pipeline for automated quantification of tumor buds in H&E whole-slide images consists of four steps (see Figure 2).

First, we determine the region of the invasive front of the tumor; second, we identify all the epithelial regions in the proximity of the invasive front via segmentation of epithelial structures at high resolution; third, we detect and characterize TB and PDC by detecting and counting neoplastic cell nuclei within the segmented epithelium compartments. Fourth, we compute the density of TBs in the entire invasive front and report the TB count as the number of TB in the hot spot. All steps of the proposed pipeline are described in the next sections.

Step 1. Find the tumor border

In the ITBCC guideline, tumor buds are be quantified within the region of the invasive front of the tumor. For this reason, the first step of our approach consists of the fully automated delineation of the tumor border, which we achieve in three steps.

First, we segment the entire whole-slide image into multiple morphological categories by applying a multi-class tissue segmentation algorithm previously developed by our group [14]. In brief, a U-Net deep learning model was trained to segment n=14 different morphological regions in the entire WSI, namely 1) normal glands, 2) low-grade dysplasia, 3) high-grade dysplasia/tumor, 4) submucosal stroma, 5) desmoplastic stroma, 6) stroma lamina propria, 7) mucus, 8) necrosis and debris, 9) lymphocytes, 10) erythrocytes, 11) adipose tissue, 12) muscle, 13) nerve, 14) background. The model was trained to operate at 10X magnification, to guarantee a fast yet accurate interpretation of the slide morphology. Additional details about this model can be found in [14].

Second, we identify the tumor bulk region by running a convex-hull algorithm on the tumor segmentation mask obtained by considering regions segmented as tumor by the multi-class algorithm. As a result, a polygon identifying the tumor bulk border is obtained.

Third, following the approach proposed in [15], we define the border as the region within 500µm from the outer side of the tumor border line.

Step 2. Segment epithelial regions

Once the tumor border region is defined, we restrict our focus to epithelial regions within the invasive front area. For this, we developed a binary segmentation model based on the U-Net architecture [16] with an EfficientNetB4 [17], to segment epithelial regions in H&E at the maximum resolution available, i.e., 0·24µm/px (40X magnification) to differentiate between very small epithelial regions, potentially containing tumor buds, and non-epithelium particles, such as activated fibroblasts to construct a binary map of all epithelial regions, which was then intersected with the invasive tumor map to generate potential candidates of TB and PDC. An overview of the training procedure can be found in Supplementary 1.

Step 3. Find tumor buds via nuclei detection

Within epithelial regions segmented inside the tumor invasive front, we identify TBs and PDCs by detecting neoplastic cell nuclei. For this purpose, we used the nuclei segmentation method based on the HoVerNet model presented by Graham et al. [18]. This network is able to segment and classify cell nuclei from six different cell types, including neoplastic cells. Here, we run HoVerNet only in epithelial regions in the invasive front, and only consider neoplastic cells within epithelial regions. In this way, we define TBs as connected components of epithelial regions containing up to four neoplastic cells, and the rest as PDC. This approach allows to 1) use HoVerNet efficiently, solely running it in selected regions, guaranteeing a fast inference time; 2) enable the analysis of the role of different number of tumor cells in TBs and their potential impact on prognosis. Step 4: Count the buds and find the hot spot.

Step 4. Find the tumor budding hotspot

The final step of our method consists in finding the hotspot of tumor buds in the entire slide and output both the location and the count of TBs in the hotspot. For this, we first slide a circular region of 0·785mm^2 over every location (within the tumor border. For each location, we consider the TBs

inside the circle and associate the TB count to the central location of the circle within the invasive front. As a result, a TB density map within the invasive front is obtained (see Figure 3C).

Reader study

In order to assess the performance of the proposed framework at the level of TB detection, we involved five pathologists in a reader study and asked them to manually mark TBs within predefined hotspots. The goal was to use the results of this study to assess human performance and inter-observer variability at the specific task of TB detection and compare our CAD system with a panel of experienced pathologists, (AL, MV, RK, SO, AE), with on average 10 years of experience in the field of tumor budding in both clinical and research setting.

We asked each pathologist to manually mark each individual tumor bud in the manually predefined hotspots of each WSI in the Val-t dataset (n=15). The hotspots were selected in line with the ITBCC guidelines and contain high amount of tumor buds. Additionally, more difficult regions were also selected (i.e., including regions with high inflammation). Slides were uploaded to the Reader Study section of the grand-challenge.org platform, and pathologists were shown both the H&E slide and the corresponding CK slide side by side.

Statistical analysis

Statistical analyses for clinical validation were performed after entry in an anonymized database using R 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria [19]). Both patient cohorts were analyzed separately. Relation of TB with outcome (overall survival (OS) and disease-free survival (DFS)) was carried out using Kaplan Meier curves and Cox regression analysis. A complete overview of the patient information can be found in Table 1.

Cut-off values for TB categories needed to be adjusted because of the increased numbers in comparison with the manual counting. To create three risk categories, we used the 30- and 60 percentiles from cohort C, which was subsequently validated in cohort D. A p-value of 0·05 was considered significant.

Comparison of manual and automatic scoring was evaluated by Cox's regression analysis in combination with the Akaike Informational Criterion (AIC, based on the log-likelihood of the model). The AIC score is a number that describes how the models compare in terms of performance given a certain dataset. When comparing two methods this value helps to identify how identical the models are.

Results

The performance of the epithelium segmentation algorithm has been assessed on dataset *Val-t* using the DICE coefficient. We applied the algorithm to the 15 whole slide images and calculated the DICE scores on the manually annotated regions. A DICE coefficient of 0·86 and 0·97 was found for the epithelium and non-epithelium class respectively.

Some examples of the segmented epithelium are shown in Figure 5A.

Validation of the tumor bud detection pipeline.

To evaluate the performance of the full detection pipeline we compared the results to the manual detections made by the five observers. Because pathologists' annotations may vary, we calculated the recall, precision, and F1 scores of all observers (the algorithm considered observer 5) relative to each other in a one-vs-one fashion. The algorithm has F1 value of 0·58 at max and recall of 0·95, but precision is suboptimal (Figure 3), with on average an amount of detected TB seven times as the other observers. An example of the manual vs. computer outputs can be found in Figure 5D.

In addition we determined per hotspot how many TB were missed by the algorithm (false negative number) compared to cumulative TB annotations of all five, all five minus 1, all five minus 2, etc. observers. In this way, we found a recall percentage of 1·0, 0·90, 0·84 for objects that have been TB annotated by at least 5, 4, 3 observers respectively.

The size of tumor buds:

For each patient group, we determined the size of TB present (Figure 4) and compared this with the manual TB annotations on dataset Val-t. In the algorithm groups these was an over-representation of single-cell TB in all groups, in contrast to the manual scores, where less single-cell TB were scored. Two-cell TB (45%) and three-cell TB (28%) were more common in the manual scoring. There was no effect of TB size on outcome.

Hotspot overlap:

We visually evaluated cases to ensure that the automatically chosen hotspots were in regions that could be used for clinical assessment according to the ITBCC guidelines. We discovered that in roughly 4 out of 5 cases the top five automatically identified hotspots were on the invasive front. In the other cases, the top 5 hotspots were incorrectly located. For example, they are not located in the invasive front but on the luminal side of the WSI.

Clinical validation: Prognostic value of TB

The presence of low TB was associated with better DFS, both in the manual scoring (cohort C: HR 2·1, 95% CI 1·6-2·8) and in the automatic scoring (cohort C HR 1·8, 95% CI 1·2 – 2·6, cohort D; HR 1·3, 95% CI 1·0 – 1·6). In the multivariable analysis, after we corrected for age, sex, T-stage, N-stage, M-stage, histological grade, and perineural invasion status for cohort C, the prognostic value of TB was still retained (cohort C manual: HR 1·8, 95% CI 1·3 – 2·5, cohort C automatic HR 1·6 95% CI 1·03 - 2·4, cohort D HR 1·2 95% CI 0·9 – 1.6) (Figure 6).

Similarly, improved overall survival was observed in patients with low TB, both in the manual scoring (cohort C: HR 2·0, 95% CI 1·6 – 2·6) and in the automatic scoring (cohort C HR 2·6, 95% CI 1·7 – 3·8, cohort D; HR 1·3, 95% CI 1·1-1·5). In the multivariable analysis, after we corrected for age, sex, T-stage, N-stage, M-stage, and histological grade, the prognostic value of TB was still retained (cohort C manual: HR 1·3, 95% CI 1·0 – 1·7, cohort C automatic HR 1·5 95% CI 1·0 - 2·3, cohort D HR 1·2 95% CI 1·0 – 2·3).

To compare the manual and automatic scoring systems we used the AIC for cohort C. For both DFS and OS the AIC was slightly lower in the manual group (790 versus 806 and 1145 versus 1148).

Discussion

In this study, we used a CNN for the segmentation of epithelium in WSI in combination with a nuclei detection network to develop a new deep learning-based automated TB assessment tool for H&E stained WSI, which we compared to the ITBCC recommended method of TB assessment, performed by five GI pathologists from different countries in an initial validation setting. We show that our method is technically sound and clinically relevant. As such, our method is not yet optimized for clinical use and could be further improved for time-efficient processing of slides. One of the main improvements that can be made is with increasing the efficiency of the epithelium segmentation and nuclei detection part of the pipeline. For example, both networks are currently applied to the invasive front at the maximum image resolution, i.e., 40X. Future work will be dedicated to investigating lowering the resolution (e.g., to 20X) for some parts of the pipeline (e.g., epithelium segmentation) without reducing overall performance, which would result in a reduction of processing time

The evident benefit of automatic assessment is standardization and minimization of interobserver variability [10]. Interobserver variability for TB is considerable [5], even when single objects should be graded [20]. In the current study variation is observed between the observers for both sensitivity/recall and specificity/precision. The algorithm is in line with the two observers who have the highest degree of interobserver agreement.

The determination of the invasive front remains a point of discussion, between pathologists and consequently, by the application of the algorithm. In this study, we opted for a fully automated selection of the tumor border to reduce the possible observer variability in hotspot selection. Although in most of the cases the hotspot was in a region pathologists would use for tumor bud scoring, the automatically selected hotspot is in some cases not in a correct location. A practical approach to circumvent this issue is a visual assessment of tumor budding heatmaps (as shown in Figure 5C) to determine the adequacy of the selected hotspot and manual correction.

The ITBCC guideline defines three different budding grades based on H&E slides, which consist of 0–4 (BD1), 5–9 (BD2) and 10 or more (BD3) buds in a hotspot. Three categories are essential, since for different clinical questions different cut-off levels are required. For example, high risk in pT1 CRC is characterized by BD2 and BD3. In contrast, only BD3 is an indication for adjuvant therapy in stage II CRC. We also assessed three-tier classifications for TB (supplemental data) with similar results to the two-tiered classification (Figure 6). Formal TB score according to the guidelines is manual, for H&E slides. For IHC it is likely that other cut-off levels need to be defined, as for automatic scoring. A recent addition to the three-tiered score is the BD0 category [21], characterized by the complete absence of TB and very good outcome. Since we trained our algorithm on TB-containing slides, and might overcall TB compared to manual scoring, we cannot be sure of adequate automatic BD0 detection. Additional testing on cases in this part of the spectrum is necessary.

TB are defined based on size: they vary between 1 and 4 tumor cells. Automatic detection shows an overrepresentation of single-cell TB (45% of TB), which partly explains the difference with the total number of TB on manual scoring. Previous studies have indicated that TB detection in IHC generates approximately three to six times more TB than TB detection in H&E [22], most likely also because of higher numbers of single-cell TB. In line, semi-automatic and automatic detection methods, that are mainly based on IHC, report higher TB [10]. With the high sensitivity of our detection tool, we detect on average about seven times as much TB as the GI pathologists per hotspot. Since we annotated our H&E slides based on cytokeratin staining patters, this can be expected.

However, another part of the increased TB can be contributed to so-called pseudobudding [6] that occurs in areas with inflammation and necrosis. These areas are avoided by pathologists. Proper segmentation is necessary to overcome this problem, illustrating that we need a combined approach in which the algorithm considers TB microenvironment.

Indeed, the overcalling of TB might be responsible for the slight loss of information when comparing manual versus automatic TB assessment, as is evident from the AIC. However, our clinical validation confirms in two different cohorts with different (local) staining protocols the prognostic value of automated TB count using the proposed algorithm. TB has independent prognostic value for both DFS and OS. Larger cohorts are required to determine the prognostic value when correcting for more extensive clinicopathological parameters (e.g., MMR, extramural venous invasion).

The proposed pipeline for automated tumor bud detection opens the door to validate the role of tumor budding in different stages or clinicopathological parameters on large multi-centric cohorts. Furthermore, it might help to understand the budding phenomenon better, as we can now start to investigate the dynamics with the surrounding macro-system, the spatial distribution throughout a single or multiple section(s). This is now possible because we can identify tumor budding in larger fields compared to the current detection area of 0.785mm^2

In conclusion, TB is an independent and relevant prognostic factor in CRC, that can be assessed by automatic methods. High TB scores show that there is a low miss rate by the algorithm, but precision might be improved, by taking the TB microenvironment into consideration and ignoring parts with pseudobudding. However, the application of novel scoring methodology requires reevaluation of current definitions and cut-off values.

Ethics approval

The Ethical Committee of the Radboud University Medical Center has approved the use of anonymized data training and validating the detection model under 2015-1637. This also includes the data from the Dublin University Hospital, Bern University, and Bayreuth University Hospital as they are part of the budding consortium agreement 10602/2016-2. The approval of the Mount Sinai Hospital was obtained under REB17-0054-E. This research was performed in accordance with the Declaration of

Helsinki. During their cancer treatment patients were informed that left-over tissue material can be used for research, and at that time, they had no objections to such use.

Authors contribution

J.B.: Conceptualization; Methodology; Investigation; Data curation; Writing – original draft.

F.C.: Conceptualization; Methodology; Supervision; Writing – original draft.

S.O.: Methodology; Writing – review & editing.

A.E.: Methodology; Writing – review & editing.

H.D.: Methodology; Writing – review & editing.

R.K.: Methodology; Writing – review & editing.

K.S.: Methodology; Writing – review & editing.

F.S.: Supervision; Writing – review & editing.

A.L.: Conceptualization; Methodology; Writing – review & editing.

I.Z.: Conceptualization; Writing – review & editing.

J.L.: Conceptualization; Supervision; Writing – original draft.

I.N.: Conceptualization; Supervision; Writing – original draft.

Conflict of interest

The authors declare no potential conflicts of interest.

Funding statement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292, from the Dutch Cancer Society, project number 10602/2016-2, The Swiss Cancer Research Foundation grand number KFS-4427-02-2018, and from the Alpe dHuZes / Dutch Cancer Society Fund, grant number KUN 2014-7032.

Data Availability Statement

The data generated in this study are available from the corresponding author upon reasonable request.

Bibliography

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

2. Brierley, J. D., Gospodarowicz, M. K., & Wittekind, C. (Eds.). (2017). *TNM classification of malignant tumours*. John Wiley & Sons.

3. Lugli, A., Kirsch, R., Ajioka, Y., Bosman, F. et al. (2017). Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. *Modern pathology, 30(9), 1299–1311*.

4. Shivji, S., Conner, J. R., Barresi, V., & Kirsch, R. (2020). Poorly differentiated clusters in colorectal cancer: a current review and implications for future practice. *Histopathology*, *77*(3), 351-368.

5. Lugli, A., Zlobec, I., Berger, M., Kirsch, R., & Nagtegaal, I. (2021). Tumour budding in solid cancers. *Nature Reviews Clinical Oncology, 18(2), 101–115.*

6. Haddad, T., Lugli, A., Aherne, S et al. (2021). Improving tumor budding reporting in colorectal cancer: a Delphi consensus study. *Virchows Archiv, 479(3), 459–469.*

7. Litjens, G., Ciompi, F., & Laak, J. (2022). A Decade of GigaScience: The Challenges of Gigapixel Pathology Images. *GigaScience, 11.*

8. Gertych, A., Swiderska-Chadaj, Z., Ma et al. (2019). Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific reports, 9(1), 1–12.*

9. Zlobec, I., & Lugli, A. (2010). Epithelial mesenchymal transition and tumor budding in aggressive colorectal cancer: tumor budding as oncotarget. *Oncotarget, 1(7), 651.*

10. Studer, L., Blank, A., Bokhorst, J.M. et al. (2021). Taking tumour budding to the next frontier—a post International Tumour Budding Consensus Conference (ITBCC) 2016 review. *Histopathology, 78(4), 476–484.*

11. Liu, S., Zhang, Y., Ju, Y. et al. (2021). Establishment and clinical application of an artificial intelligence diagnostic platform for identifying rectal cancer tumor budding. *Frontiers in Oncology, 11, 626626.*

12. Pai, R., Hartman, D., Schaeffer, D. et al. (2021). Development and initial validation of a deep learning algorithm to quantify histological features in colorectal carcinoma including tumour budding/poorly differentiated clusters. *Histopathology, 79(3), 391–405.*

13. Brand, M., Hoevenaars, B., Sigmans, J. et al. (2014). Sequential immunohistochemistry: a promising new tool for the pathology laboratory. *Histopathology, 65(5), 651–657.*

14. Bokhorst, J.M., Nagtegaal, I., Fraggetta, F. et al. (2021). Automated risk classification of colon biopsies based on semantic segmentation of histopathology images. *arXiv preprint arXiv:2109.07892*.

15. Galon, J., Marincola, F., Angell, H. et al. (2012). Cancer classification using the Immunoscore: a worldwide task force. *Journal of translational medicine, 10(1), 1–10.*

16. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241).

17. Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

18. Graham, S., Vu, Q., Raza, S. et al. (2019). Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis, 58, 101563.*

19. Team, R. (2013). R Core Team: A language and environment for statistical computing R Foundation for Statistical Computing. *Vienna, Austria*.

20. Bokhorst, J., Blank, A., Lugli, A. et al. Assessment of individual tumor buds using keratin immunohistochemistry: moderate interobserver agreement suggests a role for machine learning. *Modern pathology, 33(5), 825–833.*
21. Zlobec, I., Bächli, M., Galuppini, F., Berger, M., Dawson, H., Nagtegaal, I., & Lugli, A. (2021). Refining the ITBCC tumor budding scoring system with a "zero-budding" category in colorectal cancer. *Virchows Archiv, 479(6), 1085–1090.*

22. Fisher, N., Loughrey, M., Coleman, H., Gelbard, M., Bankhead, P., & Dunne, P. (2021). Development of a semi-automated method for tumor budding assessment in colorectal cancer and comparison with manual methods. *bioRxiv*.

Figure and table legends

Figure 1. Schematic overview of model development. A) Within corresponding H&E stained and CK-stained images manual regions of interest (ROI) were selected. Within the ROIs, all cytokeratin (CK) positive cells were annotated as either Bud or Tumor in the CK stained image. These were subsequently checked in the Hematoxylin and Eosin (H&E) stained tissue and removed if not visible or modified to follow border in H&E. B) Manual annotations were used to train the deep learning network, after which the network was applied to the train and validation set to identify hard-negatives and false-positives. In the next training session, we sampled more of these difficult objects.

Figure 2. Overview of the workflow of the algorithm. First, the input whole slide image (WSI) is segmented into fourteen tissue types. Based on the detected tumor, a 500µm invasive front border is drawn on both sides of the edge of the tumor. Within the invasive front, we identify each nucleus and segment all epithelium. After combining the two outputs we can discard all objects that have more than 4 nuclei. Based on the detected buds we can create a density map to show the regions that have the most TB.

Figure 3. F1 – scores (A) Precision (B), Recall (C), in a one-vs-one setup. For every score, one observer was set as a reference and compared to the other observers + algorithm.

Figure 4. The absolute number of 1/2/3/4 cell tumor buds found by the algorithm per patient risk group in cohorts C (left) and D (right).

Figure 5. A) Example of the output of the epithelium segmentation network (left) and the combined results (right) B) Manually (5 observers) and per algorithm obtained tumor bud (TB) counts per hotspot, where manually obtained TB range is shown from min to max with median values (marked in red), boxed by 25th and 75th percentile values. Per algorithm, detected numbers are shown as blue dots. The hotspots are sorted according to the degree of spread, from low (left) to high (right). C) Example of density heatmap generated based on the automatic tumor bud detections. D) Example of technical validation results with Hematoxylin and Eosin (H&E) stained image with manual annotations (left), corresponding Immunohistochemistry image (middle) and in blue (right) the segmented epithelium.

Figure 6. Kaplan Meier curves on disease-free survival (DFS) for A) Automatic tumor bud count in Cohort C with 30/60 percentile as cut-off values, B) Manual tumor bud counts in Cohort C with cut-off values according to the International Tumor Budding Consensus Conference, C) Automatic tumor bud counts in cohort D, with cut-off values determined by the 30/60 percentile of cohort C.

Table 1. Clinicopathological characteristics of cohort C and D.

		Cohort C		Cohort D	
Age					
	<65	226	57%	425	66%
	>65	169	43%	219	34%
Sex					
	M	215	54%	339	53%
	F	180	46%	305	47%
Invasion depth					
	T1	27	7%	18	3%
	T2	74	19%	90	14%
	T3	218	55%	402	62%
	T4	76	19%	134	21%
Nodal status					
	N0	224	57%	371	58%
	N+	171	43%	273	42%
Synchronous metastases					
	M0	351	89%	618	96%
	M1	44	11%	26	4%
Perineural invasion					
	no	83	21%	n/a	n/a
	yes	312	79%	n/a	n/a
Grade					
	low grade	350	89%	462	71%
	High grade	45	11%	182	29%
Recurrence					
	no	318	81%	502	78%
	yes	77	19%	142	22%
Death					
	no	292	74%	414	64%
	yes	103	26%	230	36%