applied sciences

MDPI

Article

Unraveling a Histopathological Needle-in-Haystack Problem: Exploring the Challenges of Detecting Tumor Budding in Colorectal Carcinoma Histology

Daniel Rusche [1,2,*], Nils Englert [3], Marlen Runz [3,4], Svetlana Hetjens [5], Cord Langner [6], Timo Gaiser [7] and Cleo-Aron Weis [3,8]

[1] Institute of Pathology, University Medical Centre Mannheim, Heidelberg University, 68167 Mannheim, Germany
[2] Department of Radiation Oncology, Technical University of Munich (TUM), Klinikum Rechts der Isar, 81675 München, Germany
[3] Institute of Pathology, University Medical Hospital Heidelberg, Heidelberg University, 69120 Heidelberg, Germany; cleo-aron.weis@uni-heidelberg.de (C.-A.W.)
[4] Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty Mannheim, Heidelberg University, 68167 Mannheim, Germany
[5] Institute for Medical Statistics, University Medical Centre Mannheim, Heidelberg University, 68167 Mannheim, Germany
[6] Diagnostic and Research Institute of Pathology, Medical University of Graz, 8036 Graz, Austria
[7] Institute of Applied Pathology, 67346 Speyer, Germany
[8] Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, 69120 Heidelberg, Germany
* Correspondence: daniel.rusche@tum.de; Tel.: +49-89-4140-4501

Abstract: Background: In this study focusing on colorectal carcinoma (CRC), we address the imperative task of predicting post-surgery treatment needs by identifying crucial tumor features within whole slide images of solid tumors, analogous to locating a needle in a histological haystack. We evaluate two approaches to address this challenge using a small CRC dataset. **Methods**: First, we explore a conventional tile-level training approach, testing various data augmentation methods to mitigate the memorization effect in a noisy label setting. Second, we examine a multi-instance learning (MIL) approach at the case level, adapting data augmentation techniques to prevent over-fitting in the limited data set context. **Results**: The tile-level approach proves ineffective due to the limited number of informative image tiles per case. Conversely, the MIL approach demonstrates success for the small dataset when coupled with post-feature vector creation data augmentation techniques. In this setting, the MIL model accurately predicts nodal status corresponding to expert-based budding scores for these cases. **Conclusions**: This study incorporates data augmentation techniques into a MIL approach, highlighting the effectiveness of the MIL method in detecting predictive factors such as tumor budding, despite the constraints of a limited dataset size.

Keywords: histopathology; CRC; budding; supervised segmentation; classification

check for updates

Citation: Rusche, D.; Englert, N.; Runz, M.; Hetjens, S.; Langner, C.; Gaiser, T.; Weis, C.-A. Unraveling a Histopathological Needle-in-Haystack Problem: Exploring the Challenges of Detecting Tumor Budding in Colorectal Carcinoma Histology. *Appl. Sci.* **2024**, *14*, 949. https://doi.org/10.3390/app14020949

Academic Editor: Francesca Silvagno

Received: 1 December 2023
Revised: 10 January 2024
Accepted: 16 January 2024
Published: 22 January 2024

1. Introduction

In this study, we address the challenge of solving a histologic needle-in-a-haystack problem related to colorectal carcinomas. Colorectal carcinoma, which is among the most prevalent malignancies globally and expected to rise in incidence in the coming years, serves as the metaphorical haystack [1,2]. The metaphorical needle to find in this context is tumor budding. In addition to the Union for International Cancer Control (UICC) TNM staging system, tumor budding could show its potential as an additional prognostic factor in recent years. Tumor budding is a long-known and long scientifically discussed histological phenomenon, first described in 1949 by Imai [3,4]. Nowadays, it is understood as a 2D-histological manifestation of the epithelial-mesenchymal-transition, as it shares many biological aspects [5,6]. Of note, in 3D, it can be appreciated that some tumor buds are tentacle-like tumor protrusions, whereas others are detached tumor cell aggregates. These detached, small tumor cell aggregates are tumor buds per definition [5]. In several publications on colorectal carcinoma (CRC) (and other solid tumors), it proved useful in predicting nodal status and overall survival [7,8]. For CRC, there are at the moment two main scenarios where tumor budding has a potential influence on clinical decision-making: First, in the case of pT1 carcinoma regarding planning the surgical approach, and second in UICC stage II, where the role of adjuvant chemotherapy still needs to be clarified in detail [6]. Though a consensus on tumor budding evaluation in CRC could be achieved, there remains considerable inter- and intra-observer variability [8–10]. Such prediction uncertainties hamper the potential clinical interpretation of such measurements. Clinical decisions could only be made if the underlying scoring system poses enough reliability. To sum it up, in theory, the task of determining the budding score per slide or, rather, slide region is a good use case for applying computer vision tools for classification (regarding a budding score) or regression (regarding tumor bud numbers). In addition, even the direct prediction of the nodal status based on the histology of the primary tumor seems to be a promising task [11]. Besides sparing humans from tedious counting tasks, the most significant benefit would be that machine learning-based classification for budding status, and subsequent nodal status prediction could reduce inter-observer variability. This would make tumor budding a reliable histomorphological biomarker. However, the relevant parts of the image—the metaphoric needles—would first have to be found or defined within the large histological sections—the metaphorical haystack. Against this background, in the presented work, we first challenge the concept of tumor budding. This is achieved through model training for nodal status prediction, followed by an examination of (a) the regions within the images harboring the most pertinent information and (b) the presence of tumor budding within these identified regions. This investigation aligns with our hypothesis concerning the significance of morphological biomarkers in histology. Subsequently, (c) we assess the prognostic efficacy of the histology-based prediction and juxtapose it with established biomarkers such as grading within our dataset. This evaluation corresponds to our hypothesis positing the independence of the identified morphological biomarker, ostensibly tumor budding.

2. Materials and Methods

In the first section of the material and methods section, we describe the data set preparation. One part of this preparation is a U-NET-segmentation-based classification of image tiles. The entire methodical proceedings and processing of data is outlined in supplementary Figure S1.

2.1. Data Set Generation

2.1.1. Data Set Collection and Digitization of the CRC Data Set

Formalin-fixed, paraffin-embedded tissue blocks and histological slides, in combination with clinical data and results of expert pathologist evaluation, were obtained from the archives of the Institute of Pathology at the Medical University of Graz, Austria. These samples were part of a previously studied patient group published by Harbaum et al. [12].

The study was conducted in compliance with the Declaration of Helsinki and approved by the local ethics committee in Graz (decision 18-199 ex 06/07). We also used parts of this data set in a previous publication on tumor budding in colorectal carcinoma CRC [13]. From the originally described 381 patients [12], 177 cases for which we have received histological slides are included in this study. For these 177 cases 643 slides have been digitalized by the Tissue Bank of the National Center for Tumor Diseases (NCT) Heidelberg, Germany in accordance with the regulations of the tissue bank and the approval of the ethics committee of Heidelberg University. For the whole slide image (WSI) generation, an Aperio AT2 scanner (Leica Biosystems, Nußloch, Germany) has been used. In total, from the included 177 cases, 643 .svs files have been generated. The range of slides included in the study per case was from 1 to 4 slides per case. This variability is due to different tumor sampling per case. For some cases, only one representative part of the tumor has been processed; for others, major parts of the tumor have been embedded, which have been considered as being representative. This embedding and inclusion strategy leads, on the one hand, to as much representative tumor tissue per case as possible, avoiding, on the other hand, duplication of tumor parts. In combination with an inherent anonymized list of patients, we included information like budding and nodal status.In this work, the mentioned nodal status Nx describes the status assessed by a pathologist commonly referred to as pNx. According to at the time of the cohort selection used TNM classification (7th edition), the nodal status was defined as N0 for cases without any lymph node metastasis, N1 for one up to three affected regional lymph nodes and N2 for more than three affected regional lymph nodes [14]. However, for some predictions the target nodal status was reduced to N− or nodal positive/pN1-2 (N+). For the data set used here, there is an expert consensus-based budding score per case originally published by Harbaum et al. in 2015 [12] and subsequently used in several other publications [13,15,16]. The budding status was defined based on Satoh's model as B0 for cases without any budding foci, B1 for one up to four budding foci, B2 for five up to nine budding foci, B3 for 10–19 budding foci and B4 for more than 20 budding foci [17]. As this evaluation method is presently considered outdated, tumor budding was not scored according to the current guidelines.

2.1.2. Segmentation Data Set Preparation and U-NET Model Training

The first step of our work was to create a segmentation model for later data preprocessing. Figure S2 shows the order and method in which we processed the original hematoxilin and eosine (HE)-images to develop a U-Net model. Therefore, a segmentation data set is needed. Whole slide colorectal tissue specimens with varying pathologies were retrieved from the collection of the Institute of Pathology at the University Medical Centre Mannheim and used in a completely anonymous way. No patient information like age or gender is included. Only the histological diagnosis (normal colorectal tissue, colorectal cancer CRC, ulcerative colitis, high grade intra-epithelial neoplasia (HGIEN), low grade intra-epithelial neoplasia (LGIEN) are used. The collection and management of this data set is in accordance with the local ethics committee in Mannheim (decision 2017-806R-MA). The final dataset contained n = 29 WSIs (n = 21 for tumor, n = 1 each for high-grade and low-grade intra-epithelial neoplasia (IEN), n = 5 for ulcerative colitis, n = 1 for healthy tissue). The whole tissue sections (HE-stained) were scanned by a M8 microscope and scanner (PreciPoint GmbH, Garching b. München, Germany). The resulting WSIs are saved in the .svs format. Using QuPath (version 0.1.2 and v0.2.0) the WSIs of the segmentation data set were manually segmented (example shown in Figure S2). For this manual segmentation, 13 labels were defined, with a ratio of area per label depicted in Table S1. The raw .svs images and the label masks were automatically cropped into 1000 × 1000 pixel-sized tiles by a QuPath-script published by Peter Bankhead (https://github.com/m4ln/qupath-scripts (accessed on 30 November 2023)) and normalized regarding staining with a tool previously described by Runz et al. [18]. This yielded a data set containing pairs of tiles, each with a label mask and the corresponding HE-frame. The data set was then randomly split into three parts training (75%), validation (15%) and testing (10%). For the segmentation task, a U-Net implementation in PyTorch (version 1.3) from the Segmentation Models toolbox by Andrew Janowczyk was downloaded from GitHub [19–21]: (https://github.com/choosehappy/PytorchDigitalPathology/tree/master/segmentation_epistroma_unet (accessed on 30 November 2023)) and adapted for multi-class segmentation. A learning rate scheduler was implemented using an Adam optimizer for efficient training [22]. Against the background of heterogeneously shaped objects and variations in the saturation of HE-staining transformations of color, size and orientation were applied on randomly selected images. Class weights were used to calculate the Focal-Loss for unbalanced labels [23,24]. The final code can be seen at [25]. The model reached an accuracy of 0.72 for the validation set. This has been seen as sufficient for this project, as our application does not require a highly specified pixel-wise segmentation but focuses on a rough estimation of the tumor content per tile.

2.1.3. Separation of the CRC Data Set into Tumor-Border and Central Tumor Regions

Using the wsi-tools and the Openslide package from Mellon University et al., the WSIs from the CRC data set were split into 1000 × 1000 pixel tiles [26,27]. These tiles underwent a primary selection to sort out all tiles of unequal edge length or those containing only a few pixels of tissue. The remaining tiles underwent segmentation with the aforementioned U-Net (compare Section 2.1.2) and were split based on the segmentation results into two groups (see Table 1):

- Group *Border* contains all tiles with 50%–75% of the tile area marked as tumor by the segmentation, here defined as the border area of the tumor;
- Group *Centre* contains all tiles with ≥95% of the tile marked as tumor area, here defined as central tumor areas (see Figure S2).

To achieve a stable classification not focusing on a case's color scheme, the tiles underwent color normalization. Therefore, we used a generative adversarial network (GAN) implemented and published by Runz et al. [18], available at https://github.com/m4ln/stainTransfer_CycleGAN_pytorch (accessed on 30 November 2023).

Table 1. Distribution of tiles for classification. Segmented HE-images were filtered for tiles depicting central tumor areas and tiles containing border tumor areas. In accordance with the patient data these tiles were distributed into the corresponding categories. We defined the following two tasks: mapping of the central tumor areas to the nodal status (hereafter referred to as Ce2No) and mapping of the tumor border areas to the nodal status (hereafter Bo2No). #—number of objects; m:f—number of male:female patients; N0:N1:N2—number of cases with corresponding nodal status; B0:B1:B2:B3:B4—number of cases with corresponding budding status.

Classification	# Cases	# Tiles	m:f	Age (mean ± std)	N0:N1:N2	B0:B1:B2:B3:B4
Bo2No	178	39,600	106:72	68.7 ± 10.9	92:39:47	3:48:45:53:27
Ce2No	169	15,734	99:70	69.0 ± 10.9	90:37:42	3:47:41:51:25

2.2. Convolutional Neural Network Model-Based Classification

In the second part of the material and methods section, we describe developing convolutional neural network (CNN) based classification models for different data sets and outcome variables. We pre-processed the images as can be seen in Figure S3. The data sorting for the classification itself is shown in Figure S4.

2.2.1. Image Classification on Tile Level Using a Deep Residual Network (Residual Neural Network (ResNet))

Data Set Preparation on Tile Level without and Concerning the Tile-Case-Relation

Two different methods were used for data set generation without and with regarding the tile-case-relation: First, for each classification, we sorted the tiles for the attribute of the classification. For the classification of Border to nodal status (Bo2No), all the tiles from peripheral tumor areas were divided into two classes according to their binary nodal status denoted in the patient data file. Then we randomly split the tiles within each class into sets for training (75%), validation (15%) and testing (10%) without regard to the WSI they belong to. This led to tiles from one single WSI being present in the training set, as well as the validation and possibly testing set. In a second approach, to respect the case origin we grouped the tiles after their corresponding WSI. The groups were then directly distributed into training, validation and test set with similar rates. Thus, tiles from one WSI could end up only in the training set, or exclusively the validation set, or the test set.

ResNet Model Training

The model used was a pre-trained residual neural network Pytorch-implementation ResNet with 152 layers [21,28]. It has been pre-trained on the ImageNet data set [29,30]. Thus, the features it relies on are trained for real-world objects like plants or trees and not histological objects. It was trained using cross-entropy loss, learning rate scheduler and stochastic gradient descend optimizer as described previously in Weis et al. [31]. The target output (or clinical output) has been the nodal status (N− or N+) of the case from which the image tiles come. The other available clinical outputs, namely budding status, nodal status and progress, have not been used here. Training was performed over 50 epochs. To increase robustness, several data augmentation methods were applied randomly, like horizontal or vertical flipping or rotation. One model has each been trained for the central tiles, the border tiles and a conjoined collection of border and central tiles, hereafter called tumor tiles.

2.2.2. Image Classification on Slide Level Using a Multi-Instance Learning (MIL) Model

Data Set Preparation for the MIL-Approach

The strategy of multiple-instance learning (MIL) involves a classification approach based on cases where each image tile is denoted as an instance. In this context, the complete case is termed a bag [32–35]. In our scenario, each bag consists of a variable number of instances, which are the tiles. Typically, labels are attributed to a whole bag rather than an individual tile; in contrast to the section above. In the MIL approach used here for WSI, the individual instances are converted into a feature vector, and then the bag containing the instances is fed in tabular form to another neural network. In our project, we generated the features with two pre-trained networks (namely a Pytorch-implemented ResNet152 model [21,28]—as above—and the HistoEncoder model [36]). In contrast to the ResNet model that has been pre-trained on real-world objects in the ImageNet data set [29,30], the HistoEncoder model has been pre-trained on large amounts of prostate tissue whole slide images [36]. Thus, the latter produces features for histological objects, albeit from another tissue type. Due to the small amount of total data (less than 200 bags), we decided against co-training the feature generation. Image tiles from the tumor border as well as the central regions are included for each case. According to the file name and known location, based on the segmentation described above (tumor central regions or border regions), the tiles are saved in folders. Subsequently, the feature vectors are calculated. Therefore, the large image tiles are further decomposed into smaller or sub-tiles, for which the feature vector is calculated. For data augmentation reasons, per 4200 × 4200 pixel image tile, ten sub-tiles of 512 × 512 pixels are randomly cropped. These small, randomly cropped sub-tiles then represent the instances. The folder represents, in this context, the bag. The number of sub-tiles or instances per bag ranges from 30 to 400.

Multi-Instance Learning MIL Model Training

The model used was based on the code from Jakub Monhart [37]. The model takes the feature vectors produced by the models mentioned above stacked in a table as input. Each row contains the feature vector of one instance. The model consists of three parts: First, a non-linear neural network for data preparation. It is applied row-wise to reduce the dimensionality. Second, an aggregation function. It is applied column-wise to map the tabular data to one vector. Here, a mean function is applied, since it does not depend on a fixed row number of the input table. Third, a non-linear neural network, in particular a fully connected network, is used to map the vector produced by the aggregation function to a class label. This is a typical classification task [37,38]. The MIL model is trained with cross-entropy loss and Adam optimizer. The target output has also been the nodal status (N− or N+) per case. Standard data augmentation techniques applied to the input images to avoid over-fitting cannot be used in this setting. Typical standard methods such as rotation or color variation typically lead to similar feature vectors in a pre-trained model (as described in the Section Data Set Preparation for the MIL approach above). The model has been trained to recognize the image content despite these modifications. Therefore, we used the query bag approach described by Babenko et al. [39] to avoid over-fitting of the MIL model. In this approach, each training a defined number of instances per bag is randomly chosen. By randomly choosing only a small fraction of the instances per bag at each iteration, over-fitting can be delayed. Nonetheless, this benefit comes at the cost of encountering noisy data. Instances relevant to the task are not usually selected randomly from a bag, which can introduce further noise. Furthermore, as additional data augmentation, we add statistical noise to the table representing the bag. Namely, we add salt-and-pepper noise (SP) to each row in the table.

2.3. Statistic Evaluation of Classification Results

Besides the standard classification metrics like accuracy and F1-score, Cohens's Kappa and the Area Under the Receiver Operating Characteristics area under the receiver operating characteristics (AUROC) are used.

2.3.1. Cohen's Kappa

To determine the agreement between the machine learning model-based classification and the ground truth we used Cohen's Kappa. Therefore, we started by calculating a k by k confusion matrix with k being the respective number of classes in this classification scenario. Within the matrix an element f_{ij} defines the items that belong to class i and are classified by the model to class j. So, f_{jj} is the number of agreements between ground truth and model prediction for class j.

Then, (from [40]):

$$P_o = \frac{1}{N}\sum_{j=1}^{k} f_{jj}, \tag{1}$$

$$r_i = \sum_{j=1}^{k} f_{ij}, \forall i, \text{ and } c_j = \sum_{i=1}^{k} f_{ij}, \forall j, \tag{2}$$

$$P_e = \frac{1}{N^2}\sum_{i=1}^{k} r_i c_i, \tag{3}$$

where P_o the observed proportional agreement, r_i and c_j the row and column totals for category i and j and P_e the expected proportion of agreement by chance. The final measure of agreement is given by Equation (4).

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{4}$$

The approximate standard deviation of κ is [40]:

$$std(\kappa) = \sqrt{\frac{P_o(1-P_o)}{N(1-P_e)^2}} \tag{5}$$

As proposed by Landis and Koch [41], Kappa values 0.6–0.8 indicate a substantial agreement and >0.8 an almost perfect agreement between two different raters.

2.3.2. Area under the Receiver Operating Characteristics

The AUROC is calculated to determine the classifier's capability to distinguish between different classes. It is based on the receiver operating characteristics (ROC) curve that plots the true positive rate (TPR) against the false positive rate (FPR) in pairs for a certain threshold, in our case the predictive probability [42].

TPR and FPR are defined as:

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity \tag{7}$$

where TP is the number of true positive results, FN is the number of false negative results, FP is the number of false positive results and TN is the number of true negative results.

We got a probability value for each class for each single output from the classifier. Taking this probability value as a threshold, we then computed the above variables for the output from the complete data set. After we successfully performed these calculations on all the outputs from the data sets, we could plot the resulting sets of TPR and FPR as an ROC curve. The AUROC represents the area under the ROC curve. An AUROC of 1 represents completely separable classes and an AUROC of 0.5 describes a classifier that is just as good as chance. For calculations, we adapted Trevisan's code from [43].

3. Results

The main objective of the presented work is to train a machine learning model to find the sparsely, inhomogeneously distributed, prognostically relevant parts of a WSI) of CRC, for a given endpoint, in our case nodal status (nodal negative/pN0 (N−) and N+). This goal is targeted within the context of a small, but heterogeneous, and therefore authentic data set. In addition, two methodologically different approaches are being tried out to achieve this goal, each of which has advantages and disadvantages:

3.1. Can Supervised Classification Model Training at the Image Tile Level (Referred to as Instance-Level) Effectively Address a Needle-in-Haystack-like Problem?

Training on the instance level has the advantage of having more than 10,000 image tiles (n = 14,448 for the central tumor parts and n = 55,334 for the border regions) instead of less than 200 cases (compare A in Figure 1). This approach is adopted to mitigate the issue of insufficient data. The advantageous aspect of this training is counterbalanced by the drawback of considerable label noise. Indeed, the sections below show that only a mean of 10% of the tiles per WSI are informative regarding the nodal status (compare B in Figure 1). In essence, this represents a needle-in-the-haystack dilemma, wherein a multitude of hay pieces and a scant number of needles share identical labels. Thus, the labels are quite noisy (compare Figure 2). During training, the models typically learn features only in the early epochs which can be seen in the loss curves. When the losses of the training and validation cohort (in a non-cross-validation setting) diverge, features detection decreases and memorization sets in [44,45]. To test, if the color normalization inhibits case and, thereby, label recognition based on the staining, we run two experiments: First, we created a training and testing data set by randomly selecting image tiles from the entire data set and, thereby, possibly spreading tiles of one case over the test, training and validation set. Through this approach, the test data set still consists of tiles not utilized to create the other two data sets. However, there are tiles from a case seen during training within the test set. Within this context, the evaluation metrics of the trained models are nearly or rather conspicuously ideal. Kappa values tend to show an almost perfect agreement with values $\geq$ 0.830 for mapping central tumor to nodal status (Ce2No) and mapping tumor border to nodal status (Bo2No). The respective area under the curve (AUC)s are $\geq$0.900 also depicting a nearly ideal separation between N− and N+. In summary, these values are highly suspicious for memorization. Second, we sorted the tiles such that all tiles of one case are either in the training, the validation or the test set. Label allocation was performed identically as before. It could already be seen during training that the model performance stagnates after a few epochs. The accuracy of the training data set and the validation data set diverge from the beginning and remain at about 0.9 and 0.55, respectively (compare the first two rows for the standard approach in Table 2). This huge gap between the loss values with a considerably elevated validation loss is a common sign of over-fitting. The validation accuracy initially stays in the range [0.50; 0.55] throughout the entire training, independent of the number of epochs or the above-mentioned early stopping strategy. In summary, the clinical feature nodal status could not be identified correctly at an AUC of 0.642, which is close to a prediction by chance (compare AUC values in Table 2). Furthermore, these results show that color normalization does not prevent recognizing cases by a CNN model, presumably based on staining characteristics.

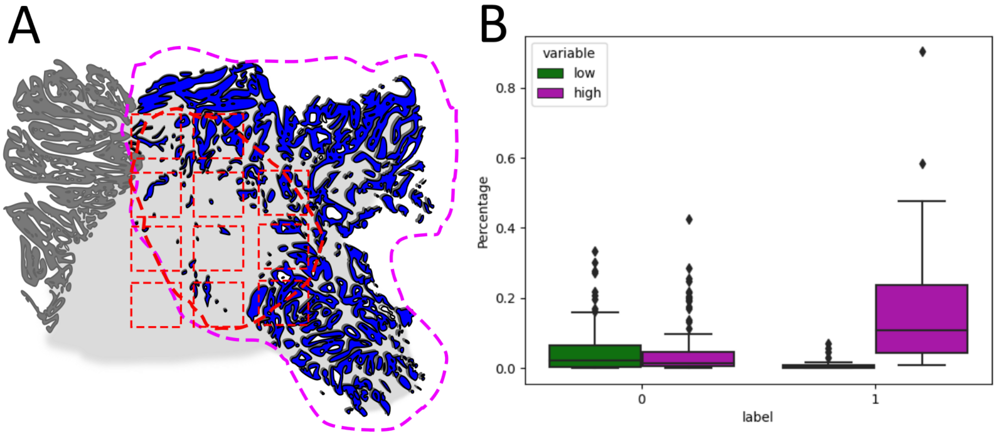

Figure 1. Region definition and distribution of informative tiles per case . (**A**) Sketch of WSI with annotated tumor region and border region. Based on a trained U-NET model the whole slide images are segmented into non-tumor (grey) and tumor (blue) tissue. The latter is subdivided based on the tumor content in each tile into tumor central (dashed line in magenta) and tumor border (dashed line in red). The tiles per region (exemplarily shown red dashed boxes) are then saved. For a graphic workflow of the tile preparations see Figure S3. (**B**) Boxplot for the frequency of meaningful instances per case. The MIL approach approach gives a class probability per instance or image tile. Per WSI, there are only a few instances with high class-probabilities. These few instances are the meaningful ones per case. The frequency of these meaningful instances per case is variable. The instances highly informative for nodal negativity are termed "low", while those with significant relevance to nodal positivity are termed "high". WSI: whole slide image; MIL: multi-instance learning.

Table 2. Statistic evaluation of the classification models. Due to their architecture standard, CNN models could perform on central or peripheral tiles separately. MIL models had to perform on the combined dataset of central and border tumor tiles, as not all methods of image augmentation could be used with this approach. AUC: area under the receiver operator characteristics curve; CNN: convolutional neuronal network; Bo2No: mapping of border tumor area tiles for their nodal status; Ce2No: mapping of central tumor area tiles for their nodal status; MIL: multiple instance learning; ResNet152: residual neural network with 152 layers.

Approach Used	Accuracy	F1-Score	Cohen's Kappa	AUC
Standard CNN Bo2No data set	0.468	0.379	0.126	0.626
Standard CNN Ce2No dataset	0.328	0.255	0.044	0.642
MIL model on ResNet152 features	0.756	0.742	0.512	0.760
MIL model on HistoEncoder features	0.683	0.648	0.389	0.794

3.2. Would Image Clustering on Image Tile-Level Work for Such a Needle-in-Haystack-like Problem?

The section above shows that standard CNN training methods do not work well with the needle-in-the-haystack problem (compare Figure 2A). At least not if all instances (hay and needles), which are assumed to be morphologically different, are given the same label.

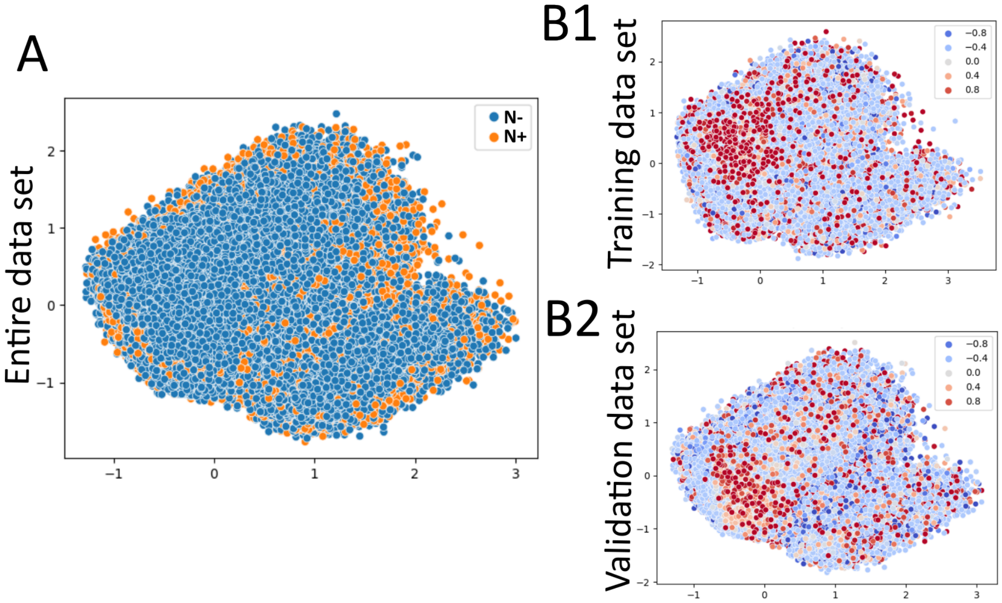

Figure 2. L-space analysis. Based on a PCA analysis of MIL results, for each instance a feature vector is plotted in 2D. (**A**) Instances from the entire data set, comprising the training and validation set, are plotted. The nodal status (nodal negative (N−) or positive (N+)) is used as a label. The silhouette score, as a clustering metric, is 0.008, which indicates a high overlap of the clusters based on the case labels. (**B**) Instances from the training set (**B1**) and the validation set (**B2**) are plotted. The label used involves the information gained concerning the nodal status. The MIL model produces a pseudo-probability per instance for the nodal status, which ranges between 0 and 1. For this plot, this range is stretched between −1 (low probability for nodal positivity and vice versa, therefore highly predictive for nodal negativity; here called "low") and +1 (high probability for nodal positivity; here called "high"). For (**B1**) the silhouette score is 0.008 and respectively for (**B2**) −0.082. PCA: principal component analysis; MIL: multi-instance learning.

This assumed morphological difference suggests cluster analysis methods as an alternative. However, plotting and subsequently analyzing the feature vectors per instance in the latent space shows no well-delineated groups. In 2-dimensional representation via Principal Component Analysis (PCA) or Uniform Manifold Approximation and Projection (UMAP), there is each time a point cloud where labeling based on the case label (nodal status) does not show well-delineated clusters with silhouette scores near 0, indicating significant overlap between the groups (Figure 2A). In conclusion, using the typical clustering methods for pre-defined groups is not promising. This also explains why the above-described CNN approaches cannot generate a clear separation. This haystack-like picture of feature vectors in the latent space appears regardless of whether models are trained completely from scratch on this data set, as in the section above, or whether pre-trained models are used. The silhouette score is with and without pre-training in the interval [−0.1; 0.1], indicating again no cluster separation but huge cluster overlap. For the pre-trained models, it also makes no difference whether the models were pre-trained on standard data sets such as ImageNet [46], whether they were trained on other circumscribed histological data sets like the glomeruli data set from a previous publication, or whether they are models for feature extraction from histological images in general like the HistoEncoder model [36]. In each case, the silhouette score is again in the range [−0.1; 0.1]. To sum it up, the information-bearing instances, the needles to stay with the image, are not morphologically distinct enough to stand out.

3.3. Does Supervised Training on Case-Level (So-Called Bag-Level) Work with a Needle-in-Haystack-like Problem?

In the methods discussed earlier, we demonstrated that addressing the data sparsity issue for cases by analyzing image tiles alone is not a viable solution. So, we need to train on the case level with, alas, only a limited data set size. After splitting the data set into training and testing sets, training a MIL model shows again that the overall data set size is too small, and the model quickly over-fits. There are approaches on the model and the data side to overcome such a data sparsity scenario: On the model side, reducing the number of trainable parameters to reduce the memorization potential is a well-known approach. However, with 10,293 trainable parameters, the used MIL model is already small in contrast to, e.g., the ResNet152 model [20,21,28] used for feature generation that has 58,149,955 trainable parameters. A reduction of trainable parameters does not avoid over-fitting in our scenario as tested for a reduction from, e.g., 31,023 to 10,293 trainable parameters (for the features produced by the pre-trained ResNet12 model).

Training a MIL model using the random query approach for data augmentation, a maximum accuracy of 0.762 and 0.721 can be achieved for the validation set, based on features from ResNet152 and HistoEncoder, respectively (compare Table 2). Per instance per bag, the N+ probability can be calculated. By doing so, image tiles can be found being informative for nodal negativity, henceforth termed "low" for being predictive for N− and, respectively, adding low risk for N+. If not, a tile might be informative for N+, henceforth termed "high" for being highly predictive for N+ (compare B in Figure 1). A mean 17.339% of the image tiles in nodal positive cases are informative for nodal positivity (N+). However, in nodal negative cases (N− cases), there are still 4.494% tiles with high probability values for nodal positivity. In summary, the predictive reliability of the used MIL model for the feature vectors per tile produced by a pre-trained network (either ResNet or HistoEncoder) is better than for models trained on the noisy tile-level (compare Section 3.1 above). Of note, in the penultimate layer, the MIL model also summarizes the given information or respectively the fed in feature vectors in its own feature vector. Visualizing the distribution of these feature vectors again in the latent space by PCA reveals a much clearer separation for the trained classes (visualized in B1 with the class labels and in B2 with the class pseudo-probabilities (in analogy to Figure 2) in Figure 3). Furthermore, a cluster map analysis (A in Figure 3) reveals that half of the features produced by the MIL model are non-informative in regard to the given task.

3.4. Does the Model Identify Tumor Budding as the Distinctive Feature?

Although the used evaluation method for the budding score is outdated (see Section 2.1.1), it is prognostically relevant concerning nodal status, as described in previous works on this data set [12].

For the data set used here, a chi-square analysis shows a strong correlation between the tumor budding score (as described above and assessed based on an outdated system) and the nodal status (p-value < 10×10^{-8}). In addition to this correlation, we tested the predictive power of the budding score per case. Therefore, the same splitting into a training and validation set as for the above-described MIL approach is used. By doing so, a logistic regression model trained to predict N− and N+ based on the budding score per case has an accuracy of 0.725 and an AUC of 0.802 for the validation set (compared to the AUC-values for the MIL approach in Table 2). Despite these prognostic capabilities, the budding score is not an independent parameter in the outdated system. For the used data set, there is a strong correlation between the expert-based tumor budding score [47] and the at this time used 3-tired tumor grading [48] per case. In the chi-square test, the p-value is also $<10^{-5}$. Thus, tumor budding is not independent of tumor grading. Testing the grading as a predictive marker for nodal status using again a logistic regression model results in an accuracy of 0.725 and an AUC of 0.690 for the validation set. In conclusion, the predictive power of grading for the nodal status seems comparable to tumor budding. In summary, the budding score in this data set is not an independent marker but harbors a predictive power. Utilizing the metrics employed (namely accuracy and AUC) to assess the predictive reliability of the tumor budding score and the previously outlined MIL approach, both exhibit comparability (compare Section 3.3 above). Besides the comparable prediction quality, both approaches show similarities in other respects so that it can be argued that the same features underlie them. From the false predicted cases in the validation data set, six cases are false in both models. Looking at the budding score values in the wrong predicted cases for both models, the MIL model and the logistic regression model, the assigned budding scores have a median of 3 (compare Figure 4. Thus, these cases show a high budding score, known as being predictive for nodal positivity, albeit both are nodal negative. The location of the relevant image tiles is also an argument that tumor budding is the underlying feature. Looking at the location of the 20% most predictive image tiles per N+ case, 84.0% of these tiles are in the border region and 16.0% are in the central tumor regions. Of the N− cases 65.3% are in the tumor border region and 34.7% are in the central regions (compare B in Figure 4). Since tumor budding is defined as taking place in the border regions, this can be used as an argument that the model at least recognizes similar features in the image.

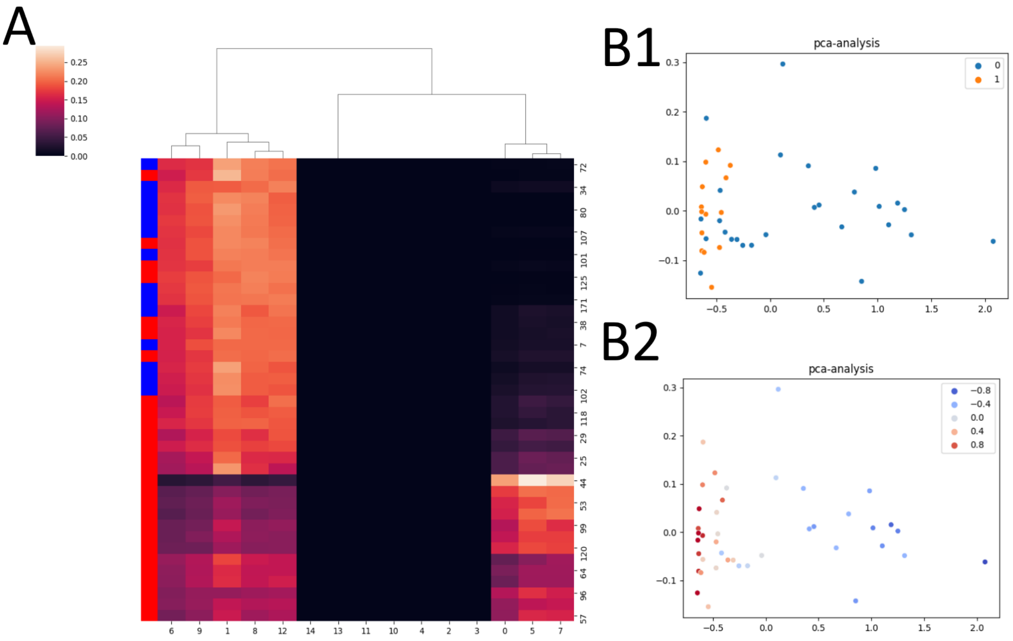

Figure 3. Clustermap analysis of the bag feature vectors produced by the MIL model. Prior to the final fully connected layer for the decision the MIL model produces a feature vector per bag based on all input instances. (**A**) The feature vector per bag is clustered against the nodal status. On the left the true status of a bag is shown with nodal negative as blue bar and nodal positive as red bar. The feature vector of one bag is split along the x-axis and different bags along the y-axis. The instances of a feature vector are depicted with color-coding representing the N+ probability (black bars equal zero). (**B1**) For the entire dataset, comprising the training and validation set, a PCA of the bag feature vectors are plotted in the latent space. The nodal status (nodal negative (0) or positive (1)) is used as a label. (**B2**) Comparable to Figure 2(B2) the pseudo-probability is shown in the L-space. Here, we compute the probability not for an instance but for whole bags for the entire dataset. Probability range is stretched between −1 (low probability for nodal positivity and vice versa, therefore highly predictive for nodal negativity; here called "low") and +1 (high probability for nodal positivity; here called "high"). MIL: multi-instance learning; PCA: principal component analysis.

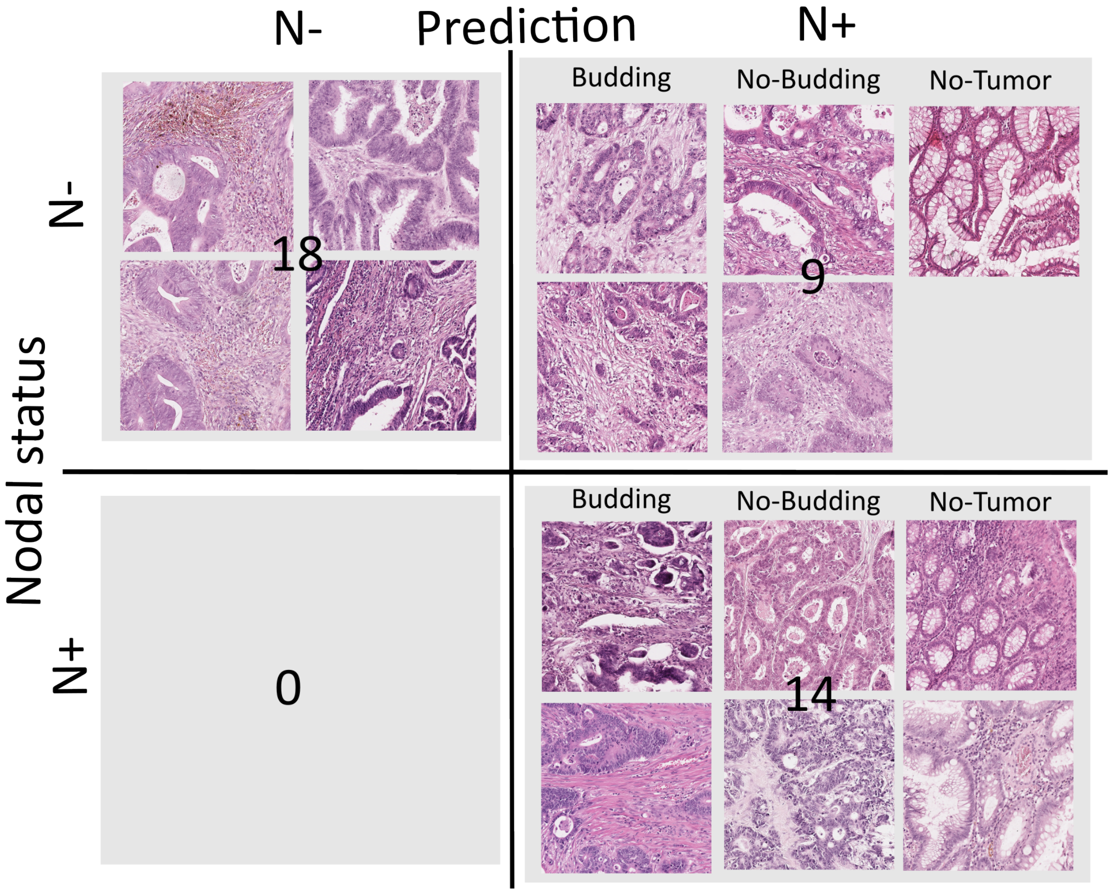

Figure 4. Confusion matrix for MIL model with exemplary tiles. The validation set (defined as 0.25 of the entire data set) comprises 31 cases or rather bags with 27 nodal negative cases (N−) and 14 nodal positive cases (N+). A total of 18 cases are correctly predicted by the MIL model as N−; and 14 cases are predicted as N+. 9 cases are false positive and 0 cases are false-negative. In summary, for the validation data set an accuracy of 0.756 and an F1-score of 0.737 is achieved. Based on the pseudo-probabilities of the network an AUC of 0.870 can be achieved. From each evaluated case, the 10 most informative tiles or the tiles with the highest significance for the respective decision are selected. AUC: area under the receiver operator characteristics curve.

4. Discussion

In this study, we elucidate the methodology for identifying a rare morphological predictive pattern analogous to locating a needle within extensive whole slide images, metaphorically representing a haystack. Specifically, we employ CRC as the haystack, predicting nodal status, all while addressing and surmounting the constraints imposed by a limited dataset size. Furthermore, we ascertain the regions within the images that harbor the most pertinent information in this context. The revelation of tumor budding in these specific areas serves to substantiate its established diagnostic significance.

4.1. Case-Level Nodal Status Prediction: A Classic Needle-in-Haystack Challenge

In the presented work, we face a typical needle-in-haystack problem with large images from which only small parts potentially contain predictive features. We can demonstrate that approximately 10% of the tiles bear significance in this context (compare Section 3.3 and Figure 2). The data set can be considered very noisy with such a low frequency of informative tiles or correct labels. Coupled with the minimal contrast between informative and non-informative feature vectors in the latent space, as assessed by close-to-zero silhouette scores (see Figure 2(B1,B2)), this complicates the training of machine learning models significantly [49]. Against this challenging background, we compared two methodologically different approaches addressing such a task: an image- and a tile-based one. First, in the tile-based approach, we convert the problem into a noisy label problem by working on the tile level. This has the advantage that such noisy label problems are well known in the literature and that there are several approaches to mitigate it [44,45,49]. Evaluating the numerous methods proposed in the literature for handling noisy label scenarios would have exceeded the scope of this paper [49]. In short, the data, the loss function, the model and/or the training settings can be adjusted to work on noisily labeled data sets [49]. From the approaches around the training settings, there are, for example, co-teaching approaches where two models are trained in parallel, updating each other [45]. For our data set, for instance, approaches based on analysis and further pre-processing are not promising, as shown in the latent space representation. In our data, as shown in Figure 2 and measured by the silhouette score, there are no distinct, pre-formed clusters to separate for in the latent space. In regard to published approaches to train a model with a noisily labeled data set, we can show that independently of countermeasures like early stopping, as the probably most easy noisy label technique, the models memorize the cases based on their staining (compare Section 3.1). Amongst the data-based approaches is the extensive usage of image augmentation techniques that modify the images before feeding into the model. One major advantage of the tile-based approach is that these standard techniques can be used without implementation effort. In regard to data augmentation techniques, however, we can demonstrate that widely recognized memorization effects in such a setting cannot be mitigated despite extensive usage of data augmentation methods, even when incorporating color normalization (compare Section 3.1 with Table 2). Second, in the case-based approach, we convert the needle-in-haystack task to a multi-instance learning setting. Approaches in this context are designed to find the informative instances (the needles in the given analogy) within a given bag of mostly non-informative instances (the hay fibers or the haystack) [32,34,50]. Our contribution to this approach is adapting data augmentation techniques. Of note, the above-mentioned standard techniques applied on the image level prior to the model do not work. On the one hand, we can show that image-based data augmentation techniques do not work for MIL approaches based on pre-trained models. On the other hand, we can show that for small data, de novo model training does not work. Thus, the data augmentation needs to take place after feature vector creation. There, however, the number of published data augmentation techniques is limited. Therefore, we adapted the query bag approach described by Babenko et al. that works on tabular data [39]. Furthermore, we add random noise to every row in the table containing the instances. By combining these data augmentation steps, we can prevent over-fitting and can show that a MIL model can be trained to generalize. Besides modifying the tabular data, in the MIL setting, there are approaches based on resampling the tabular data. This resampling is often called bag mixing [33,50]. Unlike the random sampling from bags employed in this study, bag mixing approaches proved ineffective for our data set. Under bag remixing, an early over-fitting repeatedly occurred for the data set used here. This can possibly be explained by the small overall data set size. Besides resampling, there are several publications on generating synthetic tabular data for the MIL model. This synthetic data can be generated by just interpolating between two rows. Albeit, in our case, there is no clear separation in the latent space, and the informative instances are not known a priori. Therefore, such approaches do not seem promising. And indeed, in isolated experiments , we could not demonstrate any additional benefit from the incorporation of synthetic data. Nevertheless, these were preliminary experiments, and a comprehensive assessment of all methods for generating such data exceeds the intended scope of this work [51,52].

4.2. Characterizing the Attributes of the Needles Discovered Amidst the Haystack

Based on the applied MIL-approach, we could identify the informative instances or respective image tiles within the analyzed WSI. In conclusion, we are able to find the needles within the haystack, to stick with the analogy. Looking at the histology within these seldom instances or respective needles, there are often image tiles that contain small, isolated tumor formations. These isolated tumor formations are of course well-known within the literature as tumor buds. The phenomenon is known as tumor budding, which is a manifestation of the epithelial-mesenchymal transition [5,6]. In our data set, it is found mostly in peripheral tumor regions (84.0% of the most informative tiles in regard to nodal status) and only seldom found in central tumor regions (16.0%; compare Section 3.4). Budding itself is considered to be an independent predictor of lymph node metastasis in pT1 and in stage II CRC [6,8]. It is known to mostly occur at the infiltrating tumor front [6], fitting to our findings. In comparison to the aforementioned peritumoral budding, the occurrence of intra-tumoral budding has been also linked to a lymph node metastasis with inferior scientific evidence [6,53], also fitting to our findings. In summary, by showing that amongst the most informative image parts tumor budding is present, we could argue for its (albeit well-known) diagnostic power. This is independent of discussions regarding the standardized modus operandi for measuring the tumor budding content [8]. Furthermore, it is also independent of our findings in a previous study regarding the hot spot definition for tumor budding in CRC [13]. In addition, our data set could not show an independence between tumor budding and other histological parameters. Indeed, according to our data, there is a correlation between the tumor budding score assessed by pathologists and the tumor grading. In addition, both parameters correlate comparably to the nodal status (compare Section 3.4 above). The more intriguing inquiry pertains to whether there exists an alternative morphological feature alongside tumor budding, or if tumor budding is the sole factor in question. In the literature, there are works on tumor-infiltrating lymphocytes and their prognostic capabilities. In this context, there are also image-analysis-based approaches [54,55]. Furthermore, there is evidence that the stromal reaction does contain prognostically relevant information. In this context, there are also publications from the field of digital pathology [56–58]. Both known, potentially interacting image features, namely tumor-infiltrating lymphocytes and stromal reaction, are not analyzed in our work. Based on our results, we can show by visual inspection of the most informative tiles by a board-certified pathologist that in many tiles tumor budding is present (compare Figure 4). Furthermore, in these tiles, the tumor formations seem to be less differentiated. In addition to these anecdotal points, there are also statistical considerations. The peripheral location of the most informative tiles within tumors, coupled with their comparable statistical properties (in contrast to the outdated expert-based tumor budding scoring discussed in Section 3.4) serves as evidence supporting the argument that the feature identified by the MIL model is indeed tumor budding. However, to verify this hypothesis, more experiments with known tumor budding scores per tile should be conducted. Ideally, these scores should be generated by image analysis tools, such as the one introduced by Bokhorst et al. [59].

4.3. Pros and Cons of Automated Prognostic Marker Measurement in General and for Future Routine Applications

An obvious advantage of the automatic measurement of known (as in the example of tumor budding) or unknown tissue-based markers based on HE-sections for solid tumors in general and CRC in particular is its time efficiency and inter-rater reliability. A significant challenge in establishing expert-defined markers, such as tumor budding, as a standardized prognostic tool lies in their definitions' considerable variation and, notably, in the measurement methods [6,8,60]. Moreover, the incomplete standardizations fail to determine which image sections should be analyzed, too. Various scoring systems, such as those for tumor budding, often rely on hot spots. But these systems frequently lack statistical definition or may solely represent the hottest spots [13,61]. Approaches like the MIL approach could predict, for instance, the nodal status and progress directly from the HE WSI. For a prediction of nodal status, there would be no necessity to determine budding status or other common predictors like lymphatic or submucosal invasion. These models can work, as shown here, as a black box directly producing a prediction [62,63].

Besides saving time, such black box approaches do not need expert-defined image features to rely on. This is a major advantage, as expert time is scarce, but also a major disadvantage, as an expert cannot confirm the results. In the sense of understandable artificial intelligence, a retrospective analysis—as conducted here—of the most relevant or, respectively, most informative image areas regarding their histomorphology should be conducted. In this context, MIL approaches in general are highly useful as they per architecture highlight the regions at question [33]. Furthermore, unlike humans, such models MIL approaches analyze the entire image. There is no need to define a diagnostic algorithm like for abdominal CT scans, where every part of the scan needs to be reviewed and where the so-called mirage of the first lesion needs to be actively avoided [64,65]. Furthermore, as a fully connected neural network renders the final decision, no decision rules or diagnostic reasoning algorithms need to be defined [64–66]. Of course, as such a model is a black box, the reasons for certain decisions can not be easily retrieved. To be able to understand this black box or at least make it a little more transparent, the results of such tools could be, as in this work, compared statistically to known parameters. Or a second and independent machine learning model, for example a MIL-based one, could be trained to predict the information content of every image tile fed into the first model. Apart from conducting additional experiments to elucidate the relevance of specific image areas in regard to explainability, many machine-learning approaches require additional refinement before becoming applicable for widespread use in a diagnostic setting: First, additional experiments should be conducted to check the model predictions on data sets from other institutions or so-called domains [67,68]. Second, the diagnostic accuracy of such tools, irrespective of the domain issue, necessitates validation through testing on additional data sets. For instance, understanding the false positive rate is crucial, particularly in light of potential therapeutic implications. Achieving this understanding demands extensive large-scale studies. However, in our defense, this routine application was not the focus of the proof of principle study presented here. Third, besides these content-related considerations, substantial software development is imperative to guarantee usability across diverse technological settings, among other factors. The so-called implementation gap or last mile gap can only be overcome through this [69].

4.4. Evaluating the Approach Employed in This Study Compared to Other Image Analysis-Based Methods, Focusing on Targets and Constraints

In the present era, despite human evaluation remaining the gold standard in diagnostics, there is a growing body of literature on machine learning-based image analysis in general, particularly in colorectal carcinoma. A comprehensive review of these publications is beyond the scope of this section. Readers are referred to according to reviews, such as the one by Davri et al. [70], where more than 80 publications are summarized in tabular form. The methods used in the numerous publications vary not only in methodology but also in their respective objectives. For instance, among the more than 80 publications reviewed by Davri et al. [70], only two directly concentrate on tumor budding as a readout: one by Pai et al., employing a tile-wise segmentation approach to identify diagnostically relevant patterns in HE-sections [71], and another by some authors of this paper, focusing on parts of the same data set from Graz through a tile-based classification method on IHC-sections [13]. In contrast to the methodology employed in this study, both approaches are supervised. Supervised in the sense that the diagnostically relevant structure is specified or ground truth data for tumor budding is available. In the MIL approach described here, the nodal status as the target is predefined, but not the tumor budding. Furthermore, the aforementioned study by Weis et al. considers the spatial distribution of tumor budding areas [13]. This aspect is not used in this work, as the spatial position of the image tiles is ignored (compare Section 3.3 and Figure 4). In a subsequent study, this spatial aspect could potentially be incorporated, perhaps by employing a graph-based approach, as suggested by the authors of Histocartography [72]. This small number of publications for machine learning and tumor budding is noteworthy as various diagnostic guidelines advocate tumor budding, and the evaluation of hematoxylin and eosin regarding tumor budding is still deemed the gold standard, despite being acknowledged for its limited inter- and intra-observer reliability [59]. Furthermore, among the more than 80 publications reviewed by Davri et al. [70], five publications deal with the prediction of nodal infiltration as a similar readout as our study here. Kwak et al. used a segmentation-based score for the prediction of nodal infiltration [73]. Zhao et al. and Bian et al. focus on immune infiltration in this regard, applying a tile-based evaluation with a subsequent score calculation [74,75]. Kiel et al. and Brockmoeller et al. used a score based on single-tile predictions that are subsequently summed up [11,76]. In this context, interestingly, for example, Kiel et al. utilized feature vectors from pre-trained networks to generate a tile-wise prediction, subsequently aggregated to form a case-wise prediction, a methodological similarity to the MIL approach employed in this study [11]. In summary, similar to the approach employed here, these five studies, described here, utilized supervised training aimed at a diagnostic endpoint. In this context, the lymphocytic infiltrate was identified as a pertinent feature within diagnostically relevant image areas [11,73–76]. Tumor budding was discussed as a potentially confounding, additional feature [11,76]. Irrespective of the methodological approach and the diagnostic focus, a commonality among most publications, including the MIL-based tool discussed here, is the limited integration of these tools into routine practice. A primary hurdle in this regard is that many approaches were specifically developed for a particular data set, as is evident in this case. To be applicable in routine practice, a process known as domain adaptation is essential [67]. For example, variations in HE staining protocols across institutions are so substantial that a model trained and validated in one location cannot readily apply in another. This study seeks to address this issue through the application of color normalization by using a tool from a prior publication (compare Section 2.1.3 above) [18]. Nevertheless, unlike certain aforementioned studies, the methodology presented here has not been validated on an additional independent data set, a task reserved for future investigations. However, the accompanying code is provided openly, enabling other research groups to implement and adapt this approach to their own data sets readily.

5. Conclusions, Limitations and Outlook

The work presented has two major findings: First, we can demonstrate the feasibility of combining MIL with data augmentation techniques tailored for tabular data. This integration allows for the application of machine learning to case-based annotations, exemplified in our study by the nodal status in CRC, even when working with small datasets, such as the 117 cases analyzed in this study. Second, the MIL approach, operating unsupervised at the tile level, validates the significance of tumor budding as a diagnostic morphological pattern. This confirmation is derived from observing a disproportionate presence of tumor budding in diagnostically relevant areas, aligning with existing literature on the subject. At least for the small data set here, we can show that the MIL-based approach produces similar results as the human-evaluation-based tumor budding score. Concerning limitations, the primary constraint is the small data set sourced from a single center. While we employed color normalization to address domain issues, such as staining variations, validating the tool on external datasets is imperative. Of course, such validation is not sufficient for routine use. In addition to various technical aspects (e.g., setting up an API), extensive clinical validation would also be necessary. Concerning outlook, in the future, MIL-based approaches can be used in various settings. For instance, tumor budding is not only a known prognostic factor in CRC but also in squamous cell carcinoma of the head and neck region [77]. So, future studies could easily apply the herein presented approach to tumor budding questions in other organ systems. Furthermore, as the herein-described adaptions for small data sets are not limited to tumor budding at all, the approach could be used for small data sets of, for example, uncommon tumors. For example, to stay in the head and neck region, such approaches could be used for differentiating rare spindle cell lesions of the oral cavity, where at the end, undifferentiated sarcoma types need to be separated from spindle cell carcinoma [78–81].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app14020949/s1, Figure S1: Overview of the complete workflow; Figure S2: Workflow for the segmentation of a HE-stained histological images; Table S1: Palette for segmentation labels; Figure S3: Workflow for the classification of HE-images; Figure S4: Data set preparation for the classification of HE-images.

Author Contributions: Conceptualization, C.-A.W. and T.G.; methodology, D.R. and C.-A.W.; software, D.R., M.R. and C.-A.W.; validation, D.R. and C.-A.W.; formal analysis, D.R., C.-A.W. and S.H.; investigation, D.R. and C.-A.W.; resources, C.L.; data curation, C.L. and D.R.; writing—original draft preparation, D.R., N.E. and C.-A.W.; writing—review and editing, D.R., C.-A.W., C.L., N.E. and T.G.; visualization, D.R., N.E. and C.-A.W.; supervision, T.G. and C.-A.W.; project administration, C.-A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in compliance with the Declaration of Helsinki and approved by the local ethics committee in Graz (decision 18-199 ex 06/07).

Informed Consent Statement: We obtained an ethics approval (decision 18-199 ex 06/07) from Graz, where patient consent was not deemed necessary due to the study's nature—tissue-based, retrospective, and entirely anonymous. Additionally, we secured an extra ethics approval from Mannheim (decision 2017-806R-MA), which also does not mandate patient consent.

Data Availability Statement: The code for the work is available at GitHub: https://github.com/cpheidelberg/proj_buddingCRC-MIL-pytorch (accessed on 30 November 2023).

Acknowledgments: The authors also thank the IT department staff of Medical Faculty Mannheim and especially Bohne-Lang for his support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MIL	multi-instance learning
WSI	whole slide image
AUROC	area under the receiver operating characteristics
AUC	area under the curve
Bo2No	mapping tumor border to nodal status
Ce2No	mapping central tumor to nodal status
CNN	convolutional neural network
CRC	colorectal carcinoma
FPR	false positive rate
GAN	generative adversarial network
HE	hematoxilin and eosine
HGIEN	high grade intra-epithelial neoplasia
IEN	intra-epithelial neoplasia
LGIEN	low grade intra-epithelial neoplasia
ROC	receiver operating characteristics
ResNet	residual neural network
TPR	true positive rate
WSI	whole slide image
SP	salt-and-pepper noise
N−	nodal negative/pN0
N+	nodal positive/pN1-2
UICC	Union for International Cancer Control
PCA	principal component analysis
UMAP	Uniform Manifold Approximation and Projection

References

1. Douaiher, J.; Ravipati, A.; Grams, B.; Chowdhury, S.; Alatise, O.; Are, C. Colorectal cancer—Global burden, trends, and geographical variations. *J. Surg. Oncol.* **2017**, *115*, 619–630 . [CrossRef] [PubMed]
2. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
3. Imai, T. The growth of human carcinoma: A morphological analysis. *Fukuoka Igaku Zasshi* **1954**, *45*, 102.
4. Imai, T. Histological comparison of cancer of the stomach in autopsy and operation cases. *Jpn J. Cancer Res.* **1949**, *40*, 199–201.
5. Grigore, A.D.; Jolly, M.K.; Jia, D.; Farach-Carson, M.C.; Levine, H. Tumor Budding: The Name is EMT. Partial EMT. *J. Clin. Med.* **2016**, *5*, 51. [CrossRef] [PubMed]
6. Lugli, A.; Zlobec, I.; Berger, M.D.; Kirsch, R.; Nagtegaal, I.D. Tumour budding in solid cancers. *Nat. Rev. Clin. Oncol.* **2021**, *18*, 101–115. [CrossRef]
7. Schmiegel, W.; Buchberger, B.; Follmann, M.; Graeven, U.; Heinemann, V.; Langer, T.; Nothacker, M.; Porschen, R.; Rödel, C.; Rösch, T.; et al. S3-leitlinie–kolorektales karzinom. *Z. Gastroenterol.* **2017**, *55*, 1344–1498. [CrossRef]
8. Lugli, A.; Kirsch, R.; Ajioka, Y.; Bosman, F.; Cathomas, G.; Dawson, H.; El Zimaity, H.; Fléjou, J.F.; Hansen, T.P.; Hartmann, A. Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. *Mod. Pathol.* **2017**, *30*, 1299–1311. [CrossRef]
9. Van Putten, P.G.; Hol, L.; Van Dekken, H.; Han van Krieken, J.; Van Ballegooijen, M.; Kuipers, E.J.; Van Leerdam, M.E. Inter-observer variation in the histological diagnosis of polyps in colorectal cancer screening. *Histopathology* **2011**, *58*, 974–981. [CrossRef]
10. Smits, L.J.H.; Vink-Börger, E.; Van Lijnschoten, G.; Focke-Snieders, I.; Van Der Post, R.S.; Tuynman, J.B.; Van Grieken, N.C.T.; Nagtegaal, I.D. Diagnostic variability in the histopathological assessment of advanced colorectal adenomas and early colorectal cancer in a screening population. *Histopathology* **2022**, *80*, 790–798. [CrossRef]
11. Kiehl, L.; Kuntz, S.; Höhn, J.; Jutzi, T.; Krieghoff-Henning, E.; Kather, J.N.; Holland-Letz, T.; Kopp-Schneider, A.; Chang-Claude, J.; Brobeil, A.; et al. Deep learning can predict lymph node status directly from histology in colorectal cancer. *Eur. J. Cancer* **2021**, *157*, 464–473. [CrossRef] [PubMed]
12. Harbaum, L.; Pollheimer, M.J.; Kornprat, P.; Lindtner, R.A.; Bokemeyer, C.; Langner, C. Peritumoral eosinophils predict recurrence in colorectal cancer. *Mod. Pathol.* **2015**, *28*, 403–413. [CrossRef] [PubMed]
13. Weis, C.A.; Kather, J.N.; Melchers, S.; Al-Ahmdi, H.; Pollheimer, M.J.; Langner, C.; Gaiser, T. Automatic evaluation of tumor budding in immunohistochemically stained colorectal carcinomas and correlation to clinical outcome. *Diagn. Pathol.* **2018**, *13*, 64 . [CrossRef] [PubMed]
14. Wittekind, C. *TNM: Klassifikation Maligner Tumoren*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
15. Max, N.; Harbaum, L.; Pollheimer, M.J.; Lindtner, R.A.; Kornprat, P.; Langner, C. Tumour budding with and without admixed inflammation: Two different sides of the same coin? *Br. J. Cancer* **2016**, *114*, 368–371. [CrossRef] [PubMed]
16. Betge, J.; Kornprat, P.; Pollheimer, M.J.; Lindtner, R.A.; Schlemmer, A.; Rehak, P.; Vieth, M.; Langner, C. Tumor budding is an independent predictor of outcome in AJCC/UICC stage II colorectal cancer. *Ann. Surg. Oncol.* **2012**, *19*, 3706–3712. [CrossRef] [PubMed]
17. Satoh, K.; Nimura, S.; Aoki, M.; Hamasaki, M.; Koga, K.; Iwasaki, H.; Yamashita, Y.; Kataoka, H.; Nabeshima, K. Tumor budding in colorectal carcinoma assessed by cytokeratin immunostaining and budding areas: Possible involvement of c-Met. *Cancer Sci.* **2014**, *105*, 1487–1495. [CrossRef] [PubMed]
18. Runz, M.; Rusche, D.; Schmidt, S.; Weihrauch, M.R.; Hesser, J.; Weis, C.A. Normalization of HE-stained histological images using cycle consistent generative adversarial networks. *Diagn. Pathol.* **2021**, *16*, 71. [CrossRef]
19. Janowczyk, A. Digital Pathology Segmentation using Pytorch + U-Net. In *GitHub Repository*; GitHub: San Francisco, CA, USA, 2021.
20. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in PyTorch. 2017. Available online: https://www.bibsonomy.org/bibtex/2d9d4911f0310e65b1d54ff4c13f11aad/ross_mck (accessed on 30 November 2023).
21. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
22. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
23. Dong, J. Focal Loss Improves the Model Performance on Multi-Label Image Classifications with Imbalanced Data. In Proceedings of the 2nd International Conference on Industrial Control Network and System Engineering Research, Medan, Indonesia, 3–4 September 2020; pp. 18–21. [CrossRef]
24. Mulyanto; Prakosa, S.W.; Faisal, M.; Leu, J.S. Using Optimized Focal Loss for Imbalanced Dataset on Network Intrusion Detection System. In Proceedings of the 2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring), Helsinki, Finland, 19–22 June 2022; pp. 1–7. [CrossRef]
25. Rusche, D. Segmentation and Classification of HE-Stained Colorectal Carcinoma Tissue. Available online: https://github.com/cpheidelberg/proj_buddingCRC-MIL-pytorch/tree/main/Segmentation (accessed on 30 November 2023).

26. Goode, A.; Gilbert, B.; Harkes, J. OpenSlide. Available online: https://openslide.org/ (accessed on 30 November 2023).
27. Jiang, J.; Hart, S.N. WSITools. Available online: https://github.com/smujiang/WSITools (accessed on 30 November 2023).
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
30. ImageNet. Available online: https://www.image-net.org/ (accessed on 30 November 2023).
31. Weis, C.A. Assessment of glomerular morphological patterns by deep learning. *J. Nephrol.* **2022**, *35*, 417–427. [CrossRef]
32. Ilse, M.; Tomczak, J.; Welling, M. Attention-based Deep Multiple Instance Learning. 2018. pp. 2127–2136. Available online: https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Attention-based+Deep+Multiple+Instance+Learning.+pp.+2127--2136.&btnG= (accessed on 30 November 2023).
33. Li, Z.; Zhao, W.; Shi, F.; Qi, L.; Xie, X.; Wei, Y.; Ding, Z.; Gao, Y.; Wu, S.; Liu, J.; et al. A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning. *arXiv* **2021**, arXiv:2102.03837.
34. Pevny, T.; Somol, P. Using Neural Network Formalism to Solve Multiple-Instance Problems. *arXiv* **2016**, arXiv:1609.07257.
35. Sharma, Y.; Shrivastava, A.; Ehsan, L.; Moskaluk, C.A.; Syed, S.; Brown, D.E. Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification. *arXiv* **2021**, arXiv:2103.10626.
36. Pohjonen, J. *HistoEncoder: Foundation Models for Digital Pathology*; GitHub Repository: San Francisco, CA, USA, 2023.
37. Monhart, J. Multiple Instance Learning Model Implemented in Pytorch. Available online: https://github.com/jakubmonhart/mil_pytorch (accessed on 30 November 2023).
38. CTUAvastLab/Mill.jl: Multiple Instance Learning Library Is Build on Top of Flux.jl Aimed to Prototype Flexible Multi-Instance Learning Models. Available online: https://github.com/CTUAvastLab/Mill.jl#what-is-multiple-instance-learning-mil-problem (accessed on 30 November 2023).
39. Babenko, B.; Dollár, P.; Belongie, S. *Multiple Instance Learning with Query Bags*; University of California: Oakland, CA, USA, 2006; Volume 72. [CrossRef]
40. Altman, D.G. *Practical Statistics for Medical Research*; Chapman & Hall: London, UK; CRC: London, UK, 1991.
41. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
42. Gusarova, M. Understanding AUC—ROC and Precision-Recall Curves. *Medium* . Available online: https://medium.com/@data.science.enthusiast/auc-roc-curve-ae9180eaf4f7 (accessed on 30 November 2023).
43. Trevisan, V. ROC Curve and ROC AUC. 2023. Available online: https://github.com/vinyluis/Articles#ROC%20Curve%20and%20ROC%20AUC (accessed on 30 November 2023).
44. Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.; McGuinness, K. Unsupervised label noise modeling and loss correction. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 312–321.
45. Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; Sugiyama, M. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. *arXiv* **2018**, arXiv:1804.06872.
46. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
47. Ueno, H.; Murphy, J.; Jass, J.R.; Mochizuki, H.; Talbot, I.C. Tumour 'budding' as an index to estimate the potential of aggressiveness in rectal cancer. *Histopathology* **2002**, *40*, 127–132. [CrossRef]
48. Hamilton, S. Carcinoma of the colon and rectum. In *Pathology and Genetics of Tumors of Digestive System*; IARS Press: San Francisco CA, USA, 2000.
49. Karimi, D.; Dou, H.; Warfield, S.K.; Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **2020**, *65*, 101759. [CrossRef]
50. Butke, J.; Frick, T.; Roghmann, F.; El-Mashtoly, S.F.; Gerwert, K.; Mosig, A. End-to-end multiple instance learning for whole-slide cytopathology of urothelial carcinoma. In Proceedings of the MICCAI Workshop on Computational Pathology, Virtual, 27 September 2021 ; pp. 57–68.
51. Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* **2022**, *493*, 28–45. [CrossRef]
52. Fonseca, J.; Bacao, F. Tabular and latent space synthetic data generation: A literature review. *J. Big Data* **2023**, *10*, 115. [CrossRef]
53. Lugli, A.; Karamitopoulou, E.; Zlobec, I. Tumour budding: A promising parameter in colorectal cancer. *Br. J. Cancer* **2012**, *106*, 1713–1717. [CrossRef] [PubMed]
54. Foersch, S.; Glasner, C.; Woerl, A.C.; Eckstein, M.; Wagner, D.C.; Schulz, S.; Kellers, F.; Fernandez, A.; Tserea, K.; Kloth, M.; et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat. Med.* **2023**, *29*, 430–439. [CrossRef] [PubMed]
55. Rudolf, J.; Büttner-Herold, M.; Erlenbach-Wünsch, K.; Posselt, R.; Jessberger, J.; Haderlein, M.; Hecht, M.; Hartmann, A.; Fietkau, R.; Distel, L. Regulatory T cells and cytotoxic T cells close to the epithelial–stromal interface are associated with a favorable prognosis. *OncoImmunology* **2020**, *9*, 1746149. [CrossRef] [PubMed]

56. Hacking, S.; Nasim, R.; Lee, L.; Vitkovski, T.; Thomas, R.; Shaffer, E.; Nasim, M. Whole slide imaging and colorectal carcinoma: A validation study for tumor budding and stromal differentiation. *Pathol. Res. Pract.* **2020**, *216*, 153233. [CrossRef] [PubMed]
57. Nearchou, I.P.; Kajiwara, Y.; Mochizuki, S.; Harrison, D.J.; Caie, P.D.; Ueno, H. Novel Internationally Verified Method Reports Desmoplastic Reaction as the Most Significant Prognostic Feature For Disease-specific Survival in Stage II Colorectal Cancer. *Am. J. Surg. Pathol.* **2019**, *43*, 1239–1248. [CrossRef] [PubMed]
58. Ueno, H.; Kanemitsu, Y.; Sekine, S.; Ishiguro, M.; Ito, E.; Hashiguchi, Y.; Kondo, F.; Shimazaki, H.; Mochizuki, S.; Kajiwara, Y.; et al. Desmoplastic Pattern at the Tumor Front Defines Poor-prognosis Subtypes of Colorectal Cancer. *Am. J. Surg. Pathol.* **2017**, *41*, 1506–1512. [CrossRef]
59. Bokhorst, J.M.; Nagtegaal, I.D.; Zlobec, I.; Dawson, H.; Sheahan, K.; Simmer, F.; Kirsch, R.; Vieth, M.; Lugli, A.; Laak, J.v.d.; et al. Semi-Supervised Learning to Automate Tumor Bud Detection in Cytokeratin-Stained Whole-Slide Images of Colorectal Cancer. *Cancers* **2023**, *15*, 2079. [CrossRef]
60. Lugli, A.; Vlajnic, T.; Giger, O.; Karamitopoulou, E.; Patsouris, E.S.; Peros, G.; Terracciano, L.M.; Zlobec, I. Intratumoral budding as a potential parameter of tumor progression in mismatch repair–proficient and mismatch repair–deficient colorectal cancer patients. *Hum. Pathol.* **2011**, *42*, 1833–1840. [CrossRef]
61. Koelzer, V.H.; Assarzadegan, N.; Dawson, H.; Mitrovic, B.; Grin, A.; Messenger, D.E.; Kirsch, R.; Riddell, R.H.; Lugli, A.; Zlobec, I. Cytokeratin-based assessment of tumour budding in colorectal cancer: analysis in stage II patients and prospective diagnostic experience. *J. Pathol. Clin. Res.* **2017**, *3*, 171–178. [CrossRef]
62. Bosch, S.L.; Teerenstra, S.; de Wilt, J.H.W.; Cunningham, C.; Nagtegaal, I.D. Predicting lymph node metastasis in pT1 colorectal cancer: A systematic review of risk factors providing rationale for therapy decisions. *Endoscopy* **2013**, *45*, 827–841. [CrossRef] [PubMed]
63. Pai, R.K.; Chen, Y.; Jakubowski, M.A.; Shadrach, B.L.; Plesec, T.P.; Pai, R.K. Colorectal carcinomas with submucosal invasion (pT1): Analysis of histopathological and molecular factors predicting lymph node metastasis. *Mod. Pathol.* **2017**, *30*, 113–122. [CrossRef] [PubMed]
64. Ali, M.; Evans, H.; Whitney, P.; Minhas, F.; Snead, D.R.J. Using Systemised Nomenclature of Medicine (SNOMED) codes to select digital pathology whole slide images for long-term archiving. *J. Clin. Pathol.* **2023**, *76*, 349–352. [CrossRef] [PubMed]
65. Eddy, D.M.; Clanton, C.H. The art of diagnosis: Solving the clinicopathological exercise. *N. Engl. J. Med.* **1982**, *306*, 1263–1268. [CrossRef] [PubMed]
66. Aberegg, S.K.; Callahan, S.J. Common things are common, but what is common? Incorporating probability information into differential diagnosis. *J. Eval. Clin. Pract.* **2022**, *28*, 1213–1217. [CrossRef] [PubMed]
67. Guan, H.; Liu, M.; Liu, M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* **2021**, *69*, 1173–1185. [CrossRef] [PubMed]
68. Kouw, W.M.; Loog, M. An introduction to domain adaptation and transfer learning. *arXiv* **2018**, arXiv:1812.11806.
69. Cabitza, F.; Campagner, A.; Balsano, C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. *Ann. Transl. Med.* **2020**, *8*, 501. [CrossRef]
70. Davri, A.; Birbas, E.; Kanavos, T.; Ntritsos, G.; Giannakeas, N.; Tzallas, A.T.; Batistatou, A. Deep Learning on Histopathological Images for Colorectal Cancer Diagnosis: A Systematic Review. *Diagnostics* **2022**, *12*, 837. [CrossRef]
71. Pai, R.K.; Hartman, D.; Schaeffer, D.F.; Rosty, C.; Shivji, S.; Kirsch, R.; Pai, R.K. Development and initial validation of a deep learning algorithm to quantify histological features in colorectal carcinoma including tumour budding/poorly differentiated clusters. *Histopathology* **2021**, *79*, 391–405. [CrossRef]
72. Jaume, G.; Pati, P.; Anklin, V.; Foncubierta, A.; Gabrani, M. HistoCartography: A Toolkit for Graph Analytics in Digital Pathology. In Proceedings of the MICCAI Workshop on Computational Pathology, Virtual, 27 September 2021 ; pp. 117–128.
73. Kwak, M.S.; Lee, H.H.; Yang, J.M.; Cha, J.M.; Jeon, J.W.; Yoon, J.Y.; Kim, H.I. Deep Convolutional Neural Network-Based Lymph Node Metastasis Prediction for Colon Cancer Using Histopathological Images. *Front. Oncol.* **2021**, *10*, 619803. [CrossRef] [PubMed]
74. Bian, C.; Wang, Y.; Lu, Z.; An, Y.; Wang, H.; Kong, L.; Du, Y.; Tian, J. ImmunoAIzer: A Deep Learning-Based Computational Framework to Characterize Cell Distribution and Gene Mutation in Tumor Microenvironment. *Cancers* **2021**, *13*, 1659. [CrossRef] [PubMed]
75. Zhao, M.; Yao, S.; Li, Z.; Wu, L.; Xu, Z.; Pan, X.; Lin, H.; Xu, Y.; Yang, S.; Zhang, S.; et al. The Crohn's-like lymphoid reaction density: A new artificial intelligence quantified prognostic immune index in colon cancer. *Cancer Immunol. Immunother.* **2022**, *71*, 1221–1231. [CrossRef] [PubMed]
76. Brockmoeller, S.; Echle, A.; Laleh, N.G.; Eiholm, S.; Malmstrøm, M.L.; Kuhlmann, T.P.; Levic, K.; Grabsch, H.I.; West, N.P.; Saldanha, O.L.; et al. Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer. *J. Pathol.* **2022**, *256*, 269–281. [CrossRef] [PubMed]
77. Almangush, A.; Salo, T.; Hagström, J.; Leivo, I. Tumour budding in head and neck squamous cell carcinoma—A systematic review. *Histopathology* **2014**, *65*, 587–594. [CrossRef] [PubMed]
78. Jot, K.; Nayyar, V.; Surya, V.; Mishra, D.; Sowmya, S.; Augustine, D.; Indu, M.; Haragannavar, V.C. A multicentric case study of fibroblastic and myofibroblastic oral spindle cell lesions. *J. Oral Maxillofac. Pathol.* **2023**, *27*, 629–641. [CrossRef]
79. Patel, A.M.; Choudhry, H.S.; Desai, A.D.; Shah, V.P.; Patel, P.; Eloy, J.A.; Roden, D.F.; Fang, C.H. Prognostic significance of head and neck spindle cell carcinoma. *Head Neck* **2023**, *45*, 685–696. [CrossRef]

80. Biradar, M.V.; Dantkale, S.S.; Abhange, R.S.; Kamra, H.T.; Birla, K. Spindle cell carcinoma of the tongue: A rare variant of squamous cell carcinoma. *Ecancermedicalscience* **2014**, *8*, 447. [CrossRef]
81. Vrînceanu, D.; Dumitru, M.; Ştefan, A.A.; Mogoantă, C.A.; Sajin, M. Giant pleomorphic sarcoma of the tongue base – a cured clinical case report and literature review. *Rom. J. Morphol. Embryol.* **2020**, *61*, 1323–1327. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.