Automatic Detection of Tumor Budding in Colorectal Carcinoma with Deep Learning

John-Melle Bokhorst[1,2(✉)], Lucia Rijstenberg[2], Danny Goudkade[3], Iris Nagtegaal[2], Jeroen van der Laak[1,2], and Francesco Ciompi[1,2]

[1] Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, Netherlands
john-.melle.bokhorst@radboudumc.nl
[2] Department of Pathology, Radboud University Medical Center, Nijmegen, Netherlands
[3] Department of Pathology, Maastricht University Medical Center, Maastricht, Netherlands

Abstract. Colorectal cancer patients would benefit from a valid, reliable and efficient detection of Tumor Budding (TB), as this is a proven prognostic biomarker. We explored the application of deep learning techniques to detect TB in Hematoxylin and Eosin (H&E) stained slides, and used convolutional neural networks to classify image patches as containing tumor buds, tumor glands and background. As a reference standard for training we stained slides both with H&E and immunohistochemistry (IHC), where one pathologist first annotated buds in IHC and then transferred the obtained annotations to the corresponding H&E image. We show the effectiveness of the proposed three-class approach, which allows to substantially reduce the amount of false positives, especially when combined with a hard-negative mining technique. Finally we report the results of an observer study aimed at investigating the correlation between pathologists at detecting TB in IHC and H&E.

Keywords: Deep learning · Computational pathology
Colorectal carcinoma · Tumor budding

1 Introduction

Tumor budding is defined as the presence of detached single epithelial cells or small clusters of up to 5 cells at the invasive front of colorectal cancer. It can also be found within the tumor mass, which is typically organized in irregular clusters of long stretched tumor glands. Tumor budding (TB) has received increasing attention by gastrointestinal pathologists as a promising adverse prognostic factor of lymph node and distant metastasis for colorectal carcinoma (CRC) patients. Incorporation of the phenomenon into the currently used staging system would contribute to more effective risk stratification [5]. Unfortunately, there

D. Stoyanov et al. (Eds.): COMPAY 2018/OMIA 2018, LNCS 11039, pp. 130–138, 2018.
https://doi.org/10.1007/978-3-030-00949-6_16

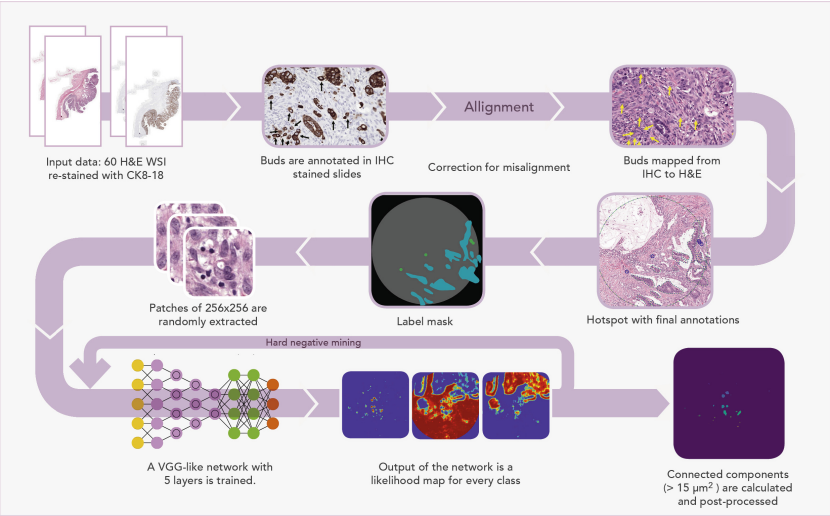


Fig. 1. Schematic overview of the proposed approach.

is no established procedure for the detection of TB so far, mainly due to the fact that there has been no reproducible method of assessment.

One of the main obstacles to the reproducibility of TB quantification has been the process of choosing the fields in which tumor budding is most intensive. In 2016, it was decided as a standard to assess the extent of TB on Hematoxylin and Eosin (H&E) in a 0.785 mm^2 hot-spot with highest TB density [5].

Microscopic identification of tumor buds in H&E by a pathologist can be difficult because of the resemblance of buds to surrounding stromal cells, fragmented glands and the concealment of buds in the setting of a peritumoral inflammatory reaction. Most studies of tumor budding have provided little detail regarding the morphologic criteria, used to include or exclude a potential bud [6]. Although tumor budding can be assessed in standard Hematoxylin and Eosin (H&E) in unproblematic cases, immunohistochemistry (IHC) with cytokeratin antibodies facilitates the detection. IHC staining highlights all epithelium and can be used for the identification of most adenocarcinomas. As IHC does not stain stromal components or tumor infiltrating lymphocytes in colon cancers, this staining can be helpful.

While manual evaluation of histological slides is still essential in clinical routine, automated image processing can provide high-throughput analysis of tumor tissue and assist the pathologist, by performing tasks such as the segmentation, classification and detection of phenomena. In recent years, Deep Learning has been leveraged to successfully address this kind of tasks. Recent developments in the field of computational pathology with Convolutional Neural Networks (CNN) are demonstrated by [1,8]. In the area of tumor budding, seminal work was done in Caie et al. [2], where immunofluorescence was employed for the

automatic detection and quantification of TB. Even though this procedure is advantageous for initial investigative purposes, immunofluorescence usually will not be applicable to clinical routine [2].

In this study we will focus on the development of a computer aided, quantitative method for detecting TB in H&E-stained CRC slides. A schematic representation of the process is given in Fig. 1. To the best of our knowledge, we are the first to pursue automated detection of TB in H&E. In order to investigate the reliability of manually obtained annotations, we conduct an observer study in which we compare the bud scores of the pathologists involved, in IHC and in H&E. We propose to build a reference standard by first detecting TB in IHC and then transferring the findings to H&E. This procedure ensures a more reliable reference standard of TB in H&E. We then use the mapped buds in the H&E slides to train a CNN for multiple class patch classification. The output of the network finally will be post-processed to obtain the bud detection.

2 Method

Materials - Data from 60 CRC patients with presence of tumor budding reported during the initial sign out were included in this study. Tissue slides were prepared from tissue blocks on which the invasive front was clearly visible, which were stained with H&E, digitized, and re-stained with CK8-18. This procedure ensured having two digitized slides of exactly the same tissue section with two different stains. Glass slides were digitized using the Pannoramic P250 Flash II scanner from 3D-Histech, at 20X magnifiction (spatial resolution of $0.24\,\mu m/px$). Following the aforementioned hot-spot driven procedure, the invasive fronts of tissue sections were visually established from low-resolution images and the one hot-spot per image was chosen. After the hot-spot was selected, buds were manually annotated by one pathologist by clicking a point in the centre of the bud and drawing a circle around it automatically, based on a pre-calculated, average bud area of $600\,\mu m^2$. In this way, an average of 5 TB per hot-spot was obtained. To obtain a reference standard, we transferred the buds annotated in the IHC slides to the H&E slides. Due to re-staining, deformations can occur, which in some cases results in misalignment in the tissue images. For this reason, we performed a semi-automatic image alignment process. This process was done by software from 3D-Histech after selection of corresponding points in the H&E and IHC slides. The transferred annotations in H&E were corrected for false positives after visual assessment by the pathologist.

Tumor buds are not only located at the invasive tumor front (peritumoral budding), but are also found within the tumor mass, namely in the stroma between the tumor glands (intratumoral budding). For this reason, having a differentiation between buds and tumor glands is beneficial, because it allows the discrimination of small tumor areas from large tumor areas, and also potentially identifies small groups of tumor cells that are part of the tumor mass, and therefore no TB. This motivated us to make annotations of tumor glands (TG) as well by delineating the tumor glands in the H&E hotspot images, which we used in the development of our method.

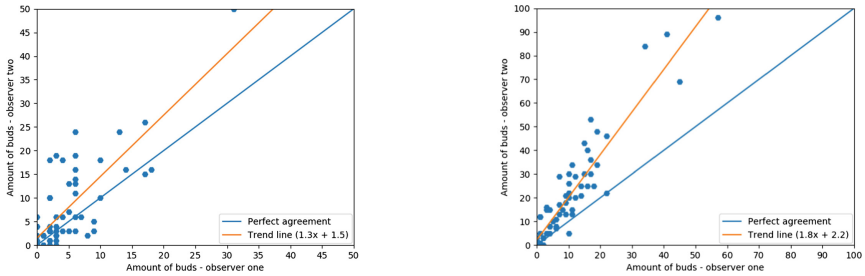
Table 1. Architecture of the CNN used in this project. MP = max-pool layer, D = dropout layer with 0.5 drop-probability, convA-B is a convolutional layer with B filters of size A × As. The last convolutional layer has C filters, where C indicates the number of classes (C = 2 or C = 3 in this paper).

conv5-32	conv5-32	MP	conv3-64	conv3-64	MP	D	conv3-128	conv3-128	MP	conv3-256	conv3-256	MP	D	conv3-512	conv3-512	avg-pool	conv4-1024	conv1-512	conv1-C	soft-max

The set of input images and corresponding annotations was randomly divided into a training (36 images, 194 buds), a validation (14 images, 73 buds in total) and a test set (10 images, 38 buds in total).

Convolutional Networks for Tumor Bud Detection - Inspired by the VGG16-net architecture [7], which was ranked at the top of ILSVRC challenge 2014, a VGG-like network was developed with two configurations: one with 2 output classes (TB versus Background) and one with 3 output classes (TB, TG, Background), as shown in Table 1. The input of both network configurations is a RGB patch of 256×256 px. This size was chosen in order to contain the surface of the area equivalent to the largest TB in our dataset, i.e., $\approx 2500\,\mu m^2$. For training purposes, class balanced patches were randomly sampled within the hot-spot. In order to sample TB patches, all pixels within the circle around the manually annotated TB location were considered. Training data were augmented by random flipping, rotating, elastic deformation, blurring, brightness (random gamma) and contrast changes. This artificially increased the number of samples and is known to optimize the network's robustness to variations in real samples of the images. Because of the relatively small amount of data and related risk of overfitting, we applied L2 regularization ($\lambda = 0.00009$), and dropout layers were added after the 2nd and 4th max-pool layer, with a drop-probability of 0.5. The densely connected layers were replaced by convolutional layers with 1024, 512 filters respectively to enable classification of arbitrary size inputs during inference. The final convolutional layer has C 1×1 filters with C representing the number of output classes. The training procedure involved optimizing the multinomial logistic regression objective (softmax), using stochastic gradient descent with Nesterov momentum. The batch size was set to 15, momentum to 0.9. We used an adaptive learning rate scheme, where the learning rate was initially set to 0.00001 and then multiplied by a factor of 0.7 after every 5 epoch if no increase in performance was observed on the validation set. The weights of the network were initialized as proposed in He et al. [3]. The networks were trained for 100 epochs.

We investigated the effect of doing hard-negative mining (HNM), applied exclusively to the output of the network with the three output classes. After classifying the training set with the trained network, we trained this network with the same settings again from scratch. In contrast to the procedure followed earlier during training, in which balanced mini-batches were used, we reduced the



(a) (b)

Fig. 2. Scatterplots of the amount of scored buds per image in (a) H&E, (b) IHC. Perfect agreement line and trend-line have been plotted.

Table 2. Corr. matrix of budding score in (left) H&E and (right) CK8-18. Classes are in line with the ITBCC 2016; Bd1 (0–4 buds), Bd2 (5–10 buds) and Bd3 (10+ buds)

Obs. 1 / Obs. 2	Bd1	Bd2	Bd3
Bd1	**31**	4	0
Bd2	3	**4**	0
Bd3	4	6	**8**

Obs. 1 / Obs. 2	Bd1	Bd2	Bd3
Bd1	**18**	0	0
Bd2	3	**2**	0
Bd3	4	8	**25**

number of patches of the third class (Background) by 50% and replaced them by hard-negatives: false positives with a probability of 0.7 (empirically determined) or higher, obtained from the likelihood maps. The number of patches of the TB and TG class were kept equal. The used 50-50 ratio was empirically observed to produce better results compared to other settings.

The output of all networks is in the form of C likelihood maps. In order to obtain the final detection of TB, we computed the center of mass of all connected components obtained after thresholding the output map of the TB class. The threshold was determined based on the validation-set and set to a probability of 0.7. Automatic detections at an euclidean distance of hand-identified buds, smaller or equal to 26 µm (which corresponds to the equavalent diameter, i.e., a single CRC tumor cell) were labeled as bud. In order to reduce TB false positives in the Tumor Glands, we applied a post-processing step to the results of the 3-class network configuration with HNM. For this purpose, we used the likelihood maps of the TG class to extract the contour of each classified glands. Bud candidates with a center or mass within the TG border and half the estimated equivalent diameter of a TB (i.e., 13 µm) were removed, as it was assumed that this implied an incomplete detachment from the tumor gland and therefore indicating a group of tumor cells that still belonged to the tumor gland itself.

In order to compute quantitative detection performance, the final detections were compared to the hand-annotated TB, in terms of F1-score and free receiver-operation characteristics (FROC) curve.

3 Observer Study

An observer study was conducted to assess inter-observer variability at TB detection on the 60 H&E and on the corresponding 60 IHC hot-spot images. In addition to the pathologist who, as described, annotated the buds in an earlier phase for the reference standard, a pathologist from another hospital was now involved. The pathologists annotated 747 buds in total in H&E. Among these detections only 143 (20%) were detected by both observers. Points closer than 26 μm from each other were considered as belonging to the same bud. On IHC a total amount of 2092 buds was identified, 570 (27%) by both pathologists. The amount of TB counted in each image by the two observers is depicted as a scatter plot in Fig. 2, where also the correlations are shown between the observers. As can be seen, the second pathologist significantly annotated more buds (x1.3 in H&E, x1.8 in IHC).

In order to get further insights on the interobserver agreement, Intraclass- and Spearman Correlation coefficients as well as Kappa values were calculated. Intraclass correlations (Two-way mixed single measures with Absolute agreement) of $r = 0.664$ (95% CI 0.442–0.798) in H&E and $r = 0.679$ (95% CI 0.233–0.847) in IHC were found. Note the relatively large 95% confidence intervals (CK range even greater than H&E range). Spearman correlation coefficients were found for H&E $r = 0.706$ and CK $r = 0.907$.

For calculation of the Kappa scores raw bud counts were classified according to ITBCC 2016 classes, see Table 2. Kappa values of 0.46 (H&E) and 0.55 (IHC) were determined.

4 Experimental Results

We evaluated the performance of the different network configurations as described in Sect. 2.2, via FROC analysis, as shown in Fig. 3. For this purpose, we first compared the performance of the networks with 2 classes and with 3 classes without HNM. As can be seen in Fig. 3 the FROC performance of the network with 3 classes is substantially better than the one with 2 classes, achieving higher sensitivity with less false positives per hot-spot, even when HNM is not used. Secondly we assessed the performance of the CNN trained with three classes when training examples were uniformly randomly sampled, and when hard-negative mining was used. It can be observed that the network with hard-negative mining gives consistently better FROC performance. Finally we compared the results of the CNN with HNM, with and without false positive reduction based on the distance from TG. As shown, the proposed post-processing technique improves the results slightly. The

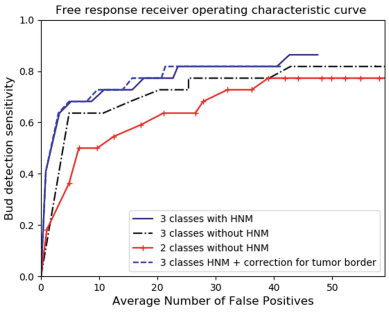
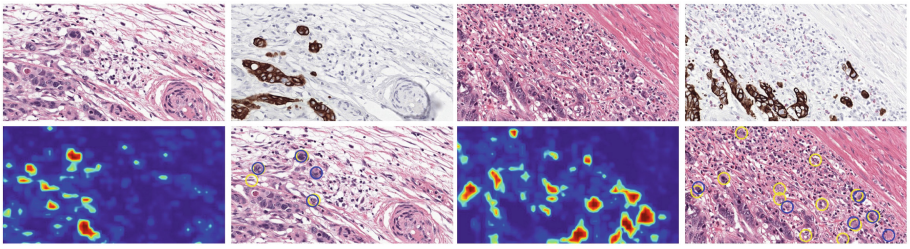


Fig. 3. FROC curves presenting the performance for all CNN's.



(a) (b)

Fig. 4. Example of details of two hot-spots (a) and (b). The original H&E stained image (top left), the CK stained image (top right), the likelihood map (bottom left) and the final output (yellow) with the manual annotations (blue) are depicted. (Color figure online)

network with 2 classes and 3 classes, reached a F1-score of 0.16 and 0.20 respectively, with recalls of 0.68 and 0.72. With HNM without post-processing a F1-score of 0.31 was reached and with HNM plus post-processing a F1-score was reached of 0.36, with a recall of 0.72 for both networks.

The likelihood maps of predicted hot-spots of the CNN with HNM and post-processing are shown in Fig. 4. A closer inspection of the accompanying maps seems to indicate that this CNN is better able to distinguish TB and stroma components from each other compared to the former CNN's, although it is also visible that it still contains false positives in the stroma area.

5 Discussion and Conclusions

In this study we explored the development of a computer aided tool for detection of TB in H&E stained images, based on convolutional neural networks. We used a VGG-like network first in a configuration with two output classes (TB and Background) and in the second instance with TG as an extra output class. We applied the method of hard-negative mining to the results of the 3-class network. The ratio for HNM was set at 50-50. We also tested with different higher ratios (more hard-negatives less Background) however, we saw that the network became less certain with regard to labeling of the class TB (lower sensitivity), a phenomenon possibly due to TB incompleteness in the reference standard. In connection with the persistent problem of the false positives in the tumor glands, we eventually applied a post-processing step to the last results. With this step, buds detected by the network in the immediate vicinity of the outline of the tumor glands were removed. However, it is clear that this procedure carries the risk that buds in the immediate vicinity of the gland contour are missed. We have analyzed the results after the post-processing step. We mainly investigated residual false positives. The analysis indicates that also the presence of larger buds (potentially poorly differentiated clusters; small clusters of >5 tumor cells) is problematic in the task to be performed, which is discriminating between TB

and TG. We therefore propose to use not only TB and TG but also PDCs as separate output class in future work.

We calculated the degree of agreement between the scores of the two pathologists using both the Spearman Correlation Coefficients and ICCs. The ICC takes into account how many buds in an image have been annotated by both pathologists, whereas the Spearman Correlation Coefficient only reflects the relationship between the number of annotated buds, and thus also gives a high correlation when the same number is scored, but not –more specifically– the same buds. Several TB investigators have included an assessment of interobserver variation in prognostic studies [4]. Based on scores from 2 or more observers, reported Kappa values for tumor budding scores range from 0.41 (moderate) to 0.938 (very good), depending on methodological factors, but also for example on the experiance of the participants. The level of agreement found in our study is moderate, although fairly in line with numbers reported by others. In view of this a better reference standard may be obtained by considering majority voting of buds detected by several pathologists in a pool of experienced observers. The reference standard can also be improved by a more precise TB annotation. In this seminal work we marked buds in the dataset by clicking and then creating an artificial outline (circle; surface equal to calculated average bud-surface). As a result, the smaller, usually single buds have been presented for analysis in conjunction with much surrounding stroma. In the future, this step could be replaced with delineating the real outlines (basement membranes) of the TB.

As our results confirm, generally a plurality of TB are found in CK, so apparently many buds in H&E are withdrawn from human perception. For this reason in our study previously marked equivalent IHC stained images were used for annotating the H&E training data set. Although this procedure will have contributed to the reliability of the reference standard in H&E, this may not have been sufficient. This conclusion is supported by our observations on the testing of several ratios for the HNM process, as we have noticed that increasing the ratio (more hard-negatives, less Background) led to a lower sensitivity. In connection with these findings, future work could focus on detection of TB in IHC staining first, whereby a procedure for better consensus on TB status between pathologists will be sought.

Acknowledgement. This project was funded by a research grant from the Dutch Cancer Society, project number 10602/2016-2. The authors would like to thank Irene Otte-Holler and Rob van de Loo for staining and scanning the WSI's.

References

1. Bejnordi, B.E., et al.: Deep learning-based assessment of tumor-associated Stroma for diagnosing breast cancer in histopathology images. In: Biomedical Imaging (ISBI 2017), pp. 929–932. IEEE (2017)
2. Caie, P.D., et al.: Quantification of tumour budding, lymphatic vessel density and invasion through image analysis in colorectal cancer. J. Transl. Med. **12**(1), 156 (2014)

3. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
4. Koelzer, V.H., Zlobec, I., Lugli, A.: Tumor budding in colorectal cancerready for diagnostic practice? Hum. Pathol. **47**(1), 4–19 (2016)
5. Lugli, A., et al.: Recommendations for reporting tumor budding in colorectal cancer based on the international tumor budding consensus conference (ITBCC) 2016. Mod. Pathol. **30**, 1299–1311 (2017)
6. Mitrovic, B., Schaeffer, D.F., Riddell, R.H., Kirsch, R.: Tumor budding in colorectal carcinoma: time to take notice. Mod. Pathol. **25**(10), 1315 (2012)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
8. Wang, D., Khosla, A., Gargeya, R., Irshad, H., Beck, A.H.: Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)