A Deep Learning Based Pipeline for Efficient Oral Cancer Screening on Whole Slide Images

Jiahao Lu[1], Nataša Sladoje[1], Christina Runow Stark[2], Eva Darai Ramqvist[3], Jan-Michaél Hirsch[4], and Joakim Lindblad[1](✉)

[1] Centre for Image Analysis, Department of IT, Uppsala University, Uppsala, Sweden
joakim@cb.uu.se

[2] Department of Orofacial Medicine at Södersjukhuset, Folktandvården Stockholms Län AB, Stockholm, Sweden

[3] Department of Clinical Pathology and Cytology, Karolinska University Hospital, Stockholm, Sweden

[4] Department of Surgical sciences, Uppsala University, Uppsala, Sweden

Abstract. Oral cancer incidence is rapidly increasing worldwide. The most important determinant factor in cancer survival is early diagnosis. To facilitate large scale screening, we propose a fully automated pipeline for oral cancer detection on whole slide cytology images. The pipeline consists of fully convolutional regression-based nucleus detection, followed by per-cell focus selection, and CNN based classification. Our novel focus selection step provides fast per-cell focus decisions at human-level accuracy. We demonstrate that the pipeline provides efficient cancer classification of whole slide cytology images, improving over previous results both in terms of accuracy and feasibility. The complete source code is made available as open source (https://github.com/MIDA-group/OralScreen).

Keywords: CNN · Whole slide imaging · Big data · Cytology · Detection · Focus selection · Classification

1 Introduction

Cancers in the oral cavity or the oropharynx are among the most common malignancies in the world [27,30]. Similar as for cervical cancer, visual inspection of brush collected samples has shown to be a practical and effective approach for early diagnosis and reduced mortality [25]. We, therefore, work towards introducing screening of high risk patients in General Dental Practice by dentists and dental hygienists. Computer assisted cytological examination is essential for feasibility of this project, due to large data and high involved costs [26].

This work is supported by: Swedish Research Council proj. 2015-05878 and 2017-04385, VINNOVA grant 2017-02447, and FTV Stockholms Län AB.

A. Campilho et al. (Eds.): ICIAR 2020, LNCS 12132, pp. 249–261, 2020.
https://doi.org/10.1007/978-3-030-50516-5_22

Whole slide imaging (WSI) refers to scanning of conventional microscopy glass slides to produce digital slides. WSI is gaining popularity among pathologists worldwide, due to its potential to improve diagnostic accuracy, increase workflow efficiency, and improve integration of images into information systems [5]. Due to the very large amount of data produced by WSI, typically generating images of around 100,000 × 100,000 pixels with up to 100,000 cells, manipulation and analysis are challenging and require special techniques. In spite of these challenges, the advantage to reproduce the traditional light microscopy experience in digital format makes WSI a very appealing choice.

Deep learning (DL) has shown to perform very well in cancer classification. An important advantage, compared to (classic) model-based approaches, is absence of need for nucleus segmentation, a difficult task typically required for otherwise subsequent feature extraction. At the same time, the large amount of data provided by WSI makes DL a natural and favorable choice. In this paper we present a complete fully automated DL based segmentation-free pipeline for oral cancer screening on WSI.

2 Background and Related Work

A number of studies suggest to use DL for classification of histology WSI samples, [1,2,18,28]. A common approach is to split tissue WSIs into smaller patches and perform analysis on the patch level. Cytological samples are, however, rather different from tissue. For tissue analysis the larger scale arrangement of cells is important and region segmentation and processing is natural. For cytology, though, the extra-cellular morphology is lost and cells are essentially individual (especially for liquid based samples); the natural unit of examination is the cell.

Cytology generally has slightly higher resolution requirements than histology; texture is very important and accurate focus is therefore essential. On the other hand, auto-focus of slide scanners works much better for tissue samples being more or less flat surfaces. In cytology, cells are partly overlapping and at different z-levels. Tools for tissue analysis rarely allow z-stacks (focus level stacks) or provide tools for handling such. In this work we present a carefully designed complete chain of processing steps for handling cytology WSIs acquired at multiple focus levels, including cell detection, per-cell focus selection, and CNN based classification.

Malignancy-associated changes (MACs) are subtle morphological changes that occur in histologically normal cells due to their proximity to a tumor. MACs have been shown to be reproducibly measured via image cytometry for numerous cancer types [29], making them potentially useful as diagnostic biomarkers. Using a random forest classifier [15] reliably detected MACs in histologically normal (normal-appearing) oropharyngeal epithelial cells located in tissue samples adjacent to a tumor and suggests to use the approach as a noninvasive means of detecting early-stage oropharyngeal tumors. Reliance on MAC enables using patient-level diagnosis for training of a cell-level classifier, where *all* cells of a patient are assigned the same label (either cancer or healthy)[32]. This hugely reduces the burden of otherwise very difficult and laborious manual annotation on a cell level.

Cell Detection: State-of-the-art object detection methods, such as the R-CNN family [7,8,23] and YOLO [20–22], have shown satisfactory performance for natural images. However, being designed for computer vision, where perspective changes the size of objects, we find them not ideal for cell detection in microscopy images. Although appealing to learn end-to-end the classification directly from the input images, s.t. the network jointly learns region of interest (RoI) selection *and* classification, for cytology WSIs this is rather impractical. The classification task is very difficult and requires tens of thousands of cells to reach top performance, while a per-cell RoI detection is much easier to train (much fewer annotated cell locations are needed), requires less detail and can be performed at lower resolution (thus faster). To jointly train localization and classification would require the (manual) localization of the full tens of thousands of cells. Our proposal, relying on patient-level annotations for the difficult classification task, reaches good performance using only around 1000 manually marked cell locations. Methods for detecting objects with various size and the bounding boxes also cost unnecessary computation, since all cell nuclei are of similar size and bounding box is not of interest in diagnosis. Further, these methods tend to not handle very large numbers of small and clustered objects very well [36].

Many DL-based methods specifically designed for the task of nucleus detection are similar to the framework summarized in [12]: first generate a probability map by sliding a binary patch classifier over the whole image, then find nuclei positions as local maxima. However, considering that WSIs are as large as 10 giga-pixels, this approach is prohibitively slow. U-Net models avoid the sliding window and reduce computation time. Detection is performed as segmentation where each nucleus is marked as a binary disk [4]. However, when images are noisy and with densely packed nuclei, the binary output mask is not ideal for further processing. We find the regression approach [16,33,34], where the network is trained to reproduce fuzzy nuclei markers, to be more appropriate.

Focus Selection: In cytological analysis, the focus level has to be selected for each nucleus individually, since different cells are at different depth. Standard tools (e.g., the microscope auto-focus) fail since they only provide a large field-of-view optimum, and often focus on clumps or other artifacts. Building on the approaches of Just Noticeable Blur (JNB) [6] and Cumulative Probability of Blur Detection (CPBD) [19], the Edge Model based Blur Metric (EMBM) [9] provides a no-reference blur metric by using a parametric edge model to detect and describe edges with both contrast and width information. It claims to achieve comparable results to the former while being faster.

Classification: Deep learning has successfully been used for different types cell classification [10] and for cervical cancer screening in particular [35]. Convolutional Neural Networks (CNNs) have shown ability to differentiate healthy and malignant cell samples [32]. Whereas the approach in [32] relies on manually selected free lying cells, our study proposes to use automatic cell detection. This allows improved performance by scaling up the available data to *all* free lying cells in each sample.

3 Materials and Methods

3.1 Data

Three sets of images of oral brush samples are used in this study. **Dataset 1** is a relatively small Pap smear dataset imaged with a standard microscope. **Dataset 2** consist of WSIs of the same glass slides as Dataset 1. **Dataset 3** consist of WSIs of liquid-based (LBC) prepared slides. All samples are collected at Dept. of Orofacial Medicine, Folktandvården Stockholms län AB. From each patient, samples were collected with a brush scraped at areas of interest in the oral cavity. Each scrape was either smeared onto a glass (Datasets 1 and 2) or placed in a liquid vial (Dataset 3). All samples were stained with standard Papanicolau stain. Dataset 3 was prepared with Hologic T5 ThinPrep Equipment and standard non-gynecologic protocol. **Dataset 1** was imaged with an Olympus BX51 bright-field microscope with a 20×, 0.75 NA lens giving a pixel size of 0.32 µm. From 10 Pap smears (10 patients), free lying cells (same as in "Oral Dataset 1" in [32]) are manually selected and 80 × 80 × 1 grayscale patches are extracted, each with one centered in-focus cell nucleus. **Dataset 2**: The same 10 slides as in Dataset 1 were imaged using a NanoZoomer S60 Digital slide scanner, 40×, 0.75 NA objective, at 11 z-offsets (±2 µm, step-size 0.4 µm) providing RGB WSIs of size 103936 × 107520 × 3, 0.23 µm/pixel. **Dataset 3** was obtained in the same way as Dataset 2, but from 12 LBC slides from 12 other patients.

Slide level annotation and reliance on MAC appears as a useful way to avoid need for large scale very difficult manual cell level annotations. Both [15] and [32] demonstrate promising results for MAC detection in histology and cytology. In our work we therefore aim to classify cells based on the patient diagnosis, i.e., all cells from a patient with diagnosed oral cancer are labeled as cancer.

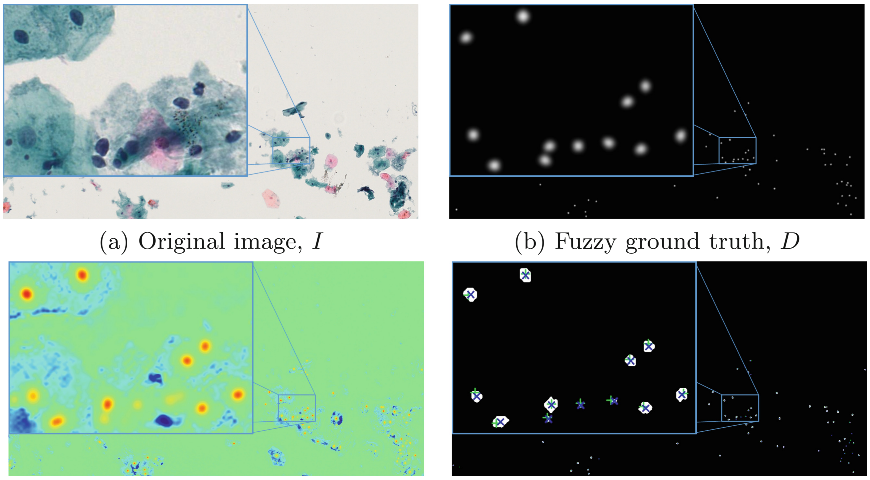
3.2 Nucleus Detection

The nucleus detection step aims to efficiently detect each individual cell nucleus in WSIs. The detection is inspired by the Fully Convolutional Regression Networks (FCRNs) approach proposed in [33] for cell counting. The main steps of the method are described below, and illustrated on an example image from Dataset 3, Fig. 1.

Training: Input is a set of RGB images $I_i, i = 1 \dots K$, and corresponding binary annotation masks B_i, where each individual nucleus is indicated by one centrally located pixel.

Each ground truth mask is dilated by a disk of radius r [4], followed by convolution with a 2D Gaussian filter of width σ. By this, a fuzzy ground truth is generated. A fully convolutional network is trained to learn a mapping between the original image I (Fig. 1a) and the corresponding "fuzzy ground truth", D (Fig. 1b). The network follows the architecture of U-Net [24] but with the final softmax replaced by a linear activation function.

Inference: A corresponding density map D' (Fig. 1c) is generated (predicted) for any given test image I. The density map D' is thresholded at a level T and centroids of the resulting blobs indicate detected nuclei locations (Fig. 1d).



(a) Original image, I

(b) Fuzzy ground truth, D

(c) Predicted density map D' (in pseudo color)

(d) Detected (blue ×) and true (green +) nuclei locations

Fig. 1. A sample image at different stages of nucleus detection (Color figure online)

3.3 Focus Selection

Slide scanners do not provide sufficiently good focus for cytological samples and a focus selection step is needed. Our proposed method utilizes N equidistant z-levels acquired of the same specimen. Traversing the z-levels, the change between consecutive images shows the largest variance at the point where the specimen moves in/out of focus. This novel focus selection approach provides a clear improvement over the Edge Model based Blur Metric (EMBM) proposed in [9].

Following the Nucleus detection step (which is performed at the central focus level, $z = 0$) we cut out a square region for each detected nucleus at all acquired focus levels. Each such cutout image is filtered with a small median filter of size $m \times m$ on each color channel to reduce noise. This gives us a set of images P_i, $i = 1, \ldots, N$, of an individual nucleus at the N consecutive z-levels. We compute the difference of neighboring focus levels, $P'_i = P_{i+1} - P_i$, $i = 1, \ldots, N-1$. The variance, σ_i^2, is computed for each difference image P'_i:

$$\sigma_i^2 = \tfrac{1}{M} \textstyle\sum_{j=1}^{M} \left(p'_{ij} - \mu_i\right)^2 \ , \text{ where } \mu_i = \tfrac{1}{M} \textstyle\sum_{j=1}^{M} p'_{ij} \ ,$$

M is the number of pixels in P'_i, and p'_{ij} is the value of pixel j in P'_i. Finally the level l corresponding to the largest σ_i^2 is selected,

$$l = \underset{i=1,\dots,N-1}{\arg\max}\ \sigma_i^2\,.$$

To determine which of the two images in the pair P'_l is in best focus, we use the EMBM method [9] as a post selection step to choose which of images P_l and P_{l+1} to use.

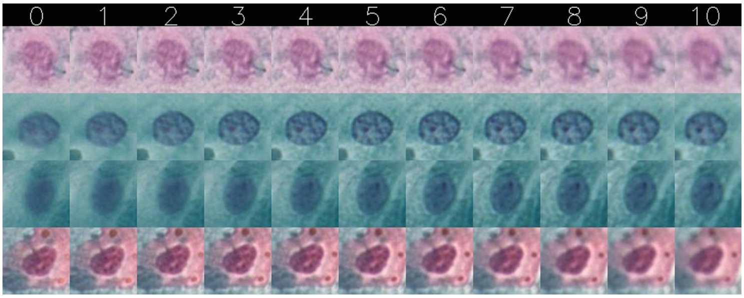


Fig. 2. Example of focus sequences for experts to annotate

3.4 Classification

The final module of the pipeline is classification of the generated nucleus patches into two classes – cancer and healthy. Following recommendation from [32], we evaluate ResNet50 [11] as a classifier. We also include the more recent DenseNet201 [13] architecture. In addition to random (Glorot-uniform) weight initialization, we also evaluate the two architectures using weights pre-trained on ImageNet.

Considering that texture information is a key feature for classification [15,31], the data is augmented without interpolation. During training, each sample is reflected with 50% probability and rotated by a random integer multiple of 90°.

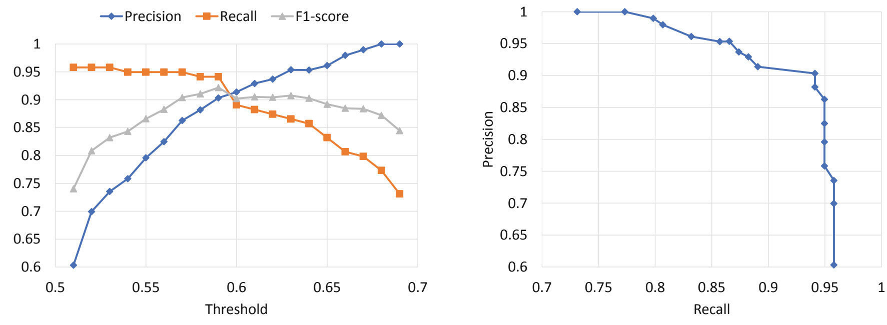
4 Experimental Setup

4.1 Nucleus Detection

The WSIs at the middle z-level ($z = 0$) are used for nucleus detection. Each WSI is split into an array of $6496 \times 3360 \times 3$ sub-images using the Open Source tool `ndpisplit` [3]. The model is trained on 12 and tested on 2 sub-images (1014 resp. 119 nuclei) from Dataset 3. The manually marked ground truth is dilated by a disk of radius $r = 15$. All images, including ground truth masks, are resized to 1024×512 pixels, using area weighted interpolation. A Gaussian filter, $\sigma = 1$, is applied to each ground truth mask providing the fuzzy ground truth D.

Each image is normalized by subtracting the mean and dividing by the standard deviation of the training set. Images are augmented by random rotation in the range ±30°, random horizontal and vertical shift within 30% of the total scale, random zoom within the range of 30% of the total size, and random horizontal and vertical flips. Nucleus detection does not need the texture details, so interpolation does not harm. To improve stability of training, batch normalization [14] is added before each activation layer. Training is performed using RMSprop with mean squared error as loss function, learning rate $\alpha = 0.001$ and decay rate $\rho = 0.9$. The model is trained with mini-batch size 1 for 100 epochs, the checkpoint with minimum training loss is used for testing.

Performance of nucleus detection is evaluated on Dataset 3. A detection is considered correct if its closest ground truth nucleus is within the cropped patch *and* that ground truth nucleus has no closer detections (s.t. one true nucleus is paired with at most one detection).



(a) Performance for different thresholds T

(b) Precision-recall curve

Fig. 3. Results of nucleus detection

4.2 Focus Selection

100 detected nuclei are randomly chosen from the two test sub-images (Dataset 3). Every nucleus is cut to an $80 \times 80 \times 3$ patch for each of the 11 z-levels. For EMBM method the contrast threshold of a salient edge is set to $c_T = 8$, following [9].

To evaluate the focus selection, 8 experts are asked to choose the best of the 11 focus-levels for each of the 100 nuclei (Fig. 2). The median of the 8 assigned labels is used as true best focus, l_{GT}. A predicted focus level l is considered accurate enough if $l \in [l_{GT} - 2, l_{GT} + 2]$.

4.3 Classification

The classification model is evaluated on Dataset 1 as a benchmark, and then on Dataset 2, to evaluate effectiveness of the nucleus detection and focus selection modules in comparison with the performance on Dataset 1. The model is also run on Dataset 3 to validate generality of the pipeline. Datasets are split on a patient level; no cell from the same patient exists in both training and test sets. On Dataset 1 and 2, three-fold validation is used, following [32]. On Dataset 3, two-fold validation is used. Our trained nucleus detector with threshold $T = 0.59$ (best-performing in Sect. 5.1) is used for Dataset 2 and 3 to generate nucleus patches. Some cells in Dataset 2 and 3 lie outside the ±2 µm imaged z-levels, and the best focus is still rather blurred. We use the EMBM to exclude the most blurred ones. Cell patches with an EMBM score < 0.03 are removed, leaving 68509 cells for Dataset 2 and 130521 for Dataset 3.

We use Adam optimizer, cross-entropy loss and parameters as suggested in [17], i.e., initial learning rate $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. 10% of the training set is randomly chosen as validation set.

When using models pre-trained on ImageNet, since the weights require three input channels, the grayscale images from Dataset 1 are duplicated into each channel. Pre-trained models are trained (fine-tuned) for 5 epochs. The learning rate is scaled by 0.4 every time the validation loss does not decrease compared to the previous epoch. The checkpoint with minimum validation loss is saved for testing.

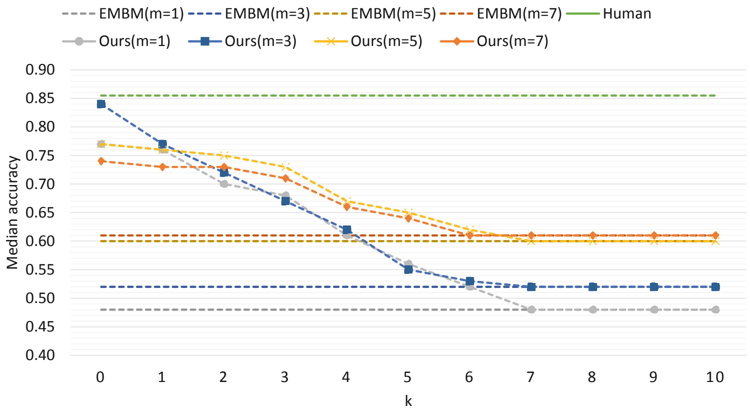


Fig. 4. Accuracy of focus selection

A slightly different training strategy is used when training from scratch. ResNet50 models are trained with mini-batch size 512 for 50 epochs on Dataset 2 and 3, and with mini-batch size 128 for 30 epochs on Dataset 1, since it contains fewer samples. Because DenseNet201 takes larger GPU memory, mini-batch sizes are set to 256 on Dataset 2 and 3. To mitigate overfitting, DenseNet201 models are trained for only 30 epochs on Dataset 2 and 3, and 20 epochs on Dataset 1. When the validation loss has not decreased for 5 epochs, the learning rate is scaled by 0.1. Training is stopped after 15 epochs of no improvement. The checkpoint with minimum validation loss is saved for testing.

5 Results and Discussion

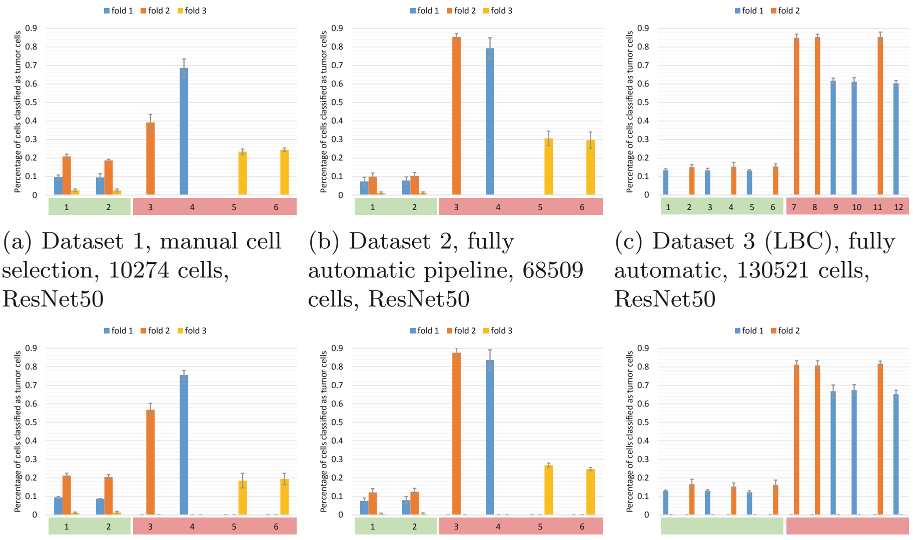
5.1 Nucleus Detection

Results of nucleus detection are presented in Fig. 3. Figure 3a shows Precision, Recall, and F1-score as the detection threshold T varies in $[0.51, 0.69]$. At $T = 0.59$, F1-score reaches 0.92, with Precision and Recall being 0.90 and 0.94 respectively. Using $T = 0.59$, 94,685 free lying nuclei are detected in Dataset 2 and 138,196 in Dataset 3.

The inference takes 0.17 s to generate a density map D' of size 1024×512 on an NVIDIA GeForce GTX 1060 Max-Q. To generate a density map of the same size based on the sliding window approach (Table 4 of [12]), takes 504 s.

5.2 Focus Selection

Performance of the focus selection is presented in Fig. 4. The "human" performance is the average of the experts, using a leave-one-out approach. We plot performance when using EMBM to select among the $2(k+1)$ levels closest to our selected pair l; for increasing k the method approaches EMBM.



(a) Dataset 1, manual cell selection, 10274 cells, ResNet50

(b) Dataset 2, fully automatic pipeline, 68509 cells, ResNet50

(c) Dataset 3 (LBC), fully automatic, 130521 cells, ResNet50

(d) Dataset 1, DenseNet201(pre-trained)

(e) Dataset 2, DenseNet201(pre-trained)

(f) Dataset 3, DenseNet201(pre-trained)

Fig. 5. Cell classification results per microscope slide; green samples (bars to the left) are healthy, red samples (bars to the right) are from cancer patients. ResNet50 is used for (a)–(c) and DenseNet201 pre-trained on ImageNet is used for (d)–(f). (Color figure online)

It can be seen that EMBM alone does not achieve satisfying performance on this task. Applying a median filter improves the performance somewhat. Our proposed method performs very well on the data and is essentially at the level of a human expert (accuracy 84% vs. 85.5%, respectively) using $k = 0$ and a 3×3 median filter.

5.3 Classification

Classification performance is presented in Table 1 and Fig. 5. The two architectures (ResNet50 and DenseNet201) perform more or less equally well. Pre-training seems to help a bit for the smaller Dataset 1, whereas for the larger Datasets 2 and 3 no essential difference is observed. Results on Dataset 2 are consistently better than on Dataset 1. This confirms effectiveness of the nucleus detection and focus selection modules; by using more nuclei (from the same samples) than those manually selected, improved performance is achieved. The results on Dataset 3 indicate that the pipeline generalizes well to liquid-based images. We also observe that our proposed pipeline is robust w.r.t. network architectures and training strategies of the classification.

In Fig. 6 we plot how classification performance decreases when nuclei are intentionally selected n focus levels away from the detected best focus. The drop in performance as we move away from the detected focus confirms the usefulness of the focus selection step.

If aggregating the cell classifications over whole microscopy slides, as show in Fig. 5, comparing Fig. 5a–5b and Fig. 5d–5e, we observe that the non-separable slides 1, 2, 5, and 6 in Dataset 1 become separable in Dataset 2. Global thresholds can be found which accurately separate the two classes of patients in both datasets processed by our pipeline.

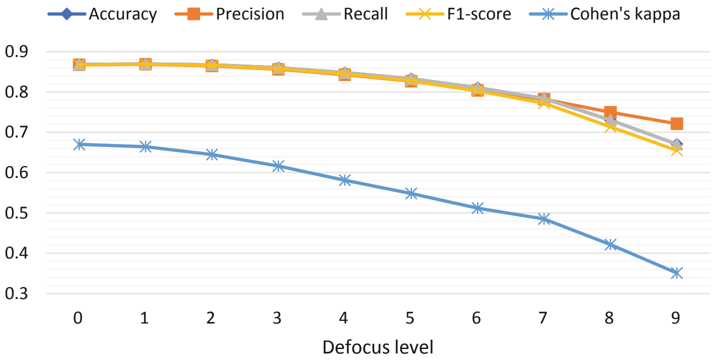


Fig. 6. The impact of defocused testset on Dataset 2, fold 1 (ResNet50)

Table 1. Classification performance. The best F1-score for each dataset is presented in bold.

Dataset	Network	Accuracy	Precision	Recall	F1-score
1	ResNet50	70.5 ± 0.5	63.1 ± 1.2	34.8 ± 1.4	44.8 ± 1.3
	ResNet50 (pre-trained)	72.0 ± 0.9	66.4 ± 2.0	37.5 ± 2.0	**48.0 ± 2.1**
	DenseNet201	70.4 ± 0.5	63.1 ± 1.8	33.8 ± 0.9	44.0 ± 0.7
	DenseNet201 (pre-trained)	70.6 ± 0.7	63.4 ± 1.6	34.3 ± 1.7	44.5 ± 1.8
2	ResNet50	74.4 ± 1.9	83.3 ± 2.9	46.3 ± 3.8	59.5 ± 3.8
	ResNet50 (pre-trained)	74.0 ± 0.1	83.9 ± 0.5	44.6 ± 0.7	58.2 ± 0.5
	DenseNet201	75.4 ± 0.8	84.3 ± 1.5	48.3 ± 1.1	**61.4 ± 1.3**
	DenseNet201 (pre-trained)	73.3 ± 0.7	81.7 ± 2.8	44.4 ± 0.3	57.5 ± 0.6
3	ResNet50	81.6 ± 0.7	71.7 ± 1.2	73.8 ± 0.9	**72.8 ± 1.0**
	ResNet50 (pre-trained)	81.3 ± 1.5	72.1 ± 3.0	71.6 ± 0.6	71.8 ± 1.8
	DenseNet201	81.3 ± 0.5	71.4 ± 0.7	73.0 ± 0.8	72.2 ± 0.7
	DenseNet201 (pre-trained)	81.5 ± 1.3	71.2 ± 2.4	74.5 ± 2.4	72.8 ± 1.9

6 Conclusion

This work presents a complete fully automated pipeline for oral cancer screening on whole slide images; source code (utilizing TensorFlow 1.14) is shared as open source. The proposed focus selection method performs at the level of a human expert and significantly outperforms EMBM. The pipeline can provide fully automatic inference for WSIs within reasonable computation time. It performs well for smears as well as liquid-based slides.

Comparing the performance on Dataset 1, using human selected nuclei and Dataset 2, using computer selected nuclei from the same microscopy slides, we conclude that the presented pipeline can reduce human workload while at the same time make the classification easier and more reliable.

References

1. Campanella, G., et al.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (2019)
2. Cruz-Roa, A., et al.: Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. Sci. Rep. **7**, 46450 (2017)
3. Deroulers, C., Ameisen, D., Badoual, M., Gerin, C., Granier, A., Lartaud, M.: Analyzing huge pathology images with open source software. Diagnostic Pathol. **8**(1), 92 (2013)
4. Falk, T., et al.: U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods **16**(1), 67–70 (2019)
5. Farahani, N., Parwani, A., Pantanowitz, L.: Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol. Lab Med. Int. **7**, 23–33 (2015)

6. Ferzli, R., Karam, L.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). IEEE Trans. Image Process. **18**(4), 717–728 (2009)
7. Girshick, R.: Fast R-CNN. arXiv:1504.08083 [cs], September 2015
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:1311.2524 [cs], October 2014
9. Guan, J., Zhang, W., Gu, J., Ren, H.: No-reference blur assessment based on edge modeling. J. Vis. Commun. Image Represent. **29**, 1–7 (2015)
10. Gupta, A., et al.: Deep learning in image cytometry: a review. Cytometry Part A **95**(4), 366–380 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv:1512.03385 [cs], December 2015
12. Höfener, H., Homeyer, A., Weiss, N., Molin, J., Lundström, C., Hahn, H.: Deep learning nuclei detection: a simple approach can deliver state-of-the-art results. Comput. Med. Imag. Graph. **70**, 43–52 (2018)
13. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. arXiv:1608.06993 [cs], January 2018
14. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 [cs] (2015)
15. Jabalee, J., et al.: Identification of malignancy-associated changes in histologically normal tumor-adjacent epithelium of patients with HPV-positive oropharyngeal cancer. Anal. Cellular Pathol. **2018**, 1–9 (2018)
16. Kainz, P., Urschler, M., Schulter, S., Wohlhart, P., Lepetit, V.: You should use regression to detect cells. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 276–283. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_33
17. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 [cs], December 2014
18. Korbar, B., et al.: Deep learning for classification of colorectal polyps on whole-slide images. J. Pathol. Inform. **8**, 30 (2017)
19. Narvekar, N.D., Karam, L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). IEEE Trans. Image Process. **20**(9), 2678–2683 (2011)
20. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of IEEE Conference on CVPR, pp. 779–788 (2016)
21. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of IEEE Conference on CVPR, pp. 7263–7271 (2017)
22. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs], April 2018
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv:1506.01497 [cs], January 2016
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) Medical Image Computing and Computer Assisted Intervention MICCAI 2015, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
25. Sankaranarayanan, R., et al.: Long term effect of visual screening on oral cancer incidence and mortality in a randomized trial in Kerala. India. Oral Oncol. **49**(4), 314–321 (2013)

26. Speight, P., et al.: Screening for oral cancer—a perspective from the global oral cancer forum. Oral Surg., Oral Med. Oral Pathol. Oral Radiol. **123**(6), 680–687 (2017)
27. Stewart, B., Wild, C.P., et al.: World cancer report 2014. Public Health (2014)
28. Teramoto, A., et al.: Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. Inform. Med. Unlocked **16**, 100205 (2019)
29. Us-Krasovec, M., et al.: Malignancy associated changes in epithelial cells of buccal mucosa: a potential cancer detection test. Anal. Quantit. Cytol. Histol. **27**(5), 254–262 (2005)
30. Warnakulasuriya, S.: Global epidemiology of oral and oropharyngeal cancer. Oral Oncol. **45**(4–5), 309–316 (2009)
31. Wetzer, E., Gay, J., Harlin, H., Lindblad, J., Sladoje, N.: When texture matters: texture-focused CNNs outperform general data augmentation and pretraining in oral cancer detection. In: Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI) (2020) forthcoming
32. Wieslander, H., et al.: Deep convolutional neural networks for detecting cellular changes due to malignancy. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 82–89. IEEE, October 2017
33. Xie, W., Noble, J., Zisserman, A.: Microscopy cell counting and detection with fully convolutional regression networks. Comput. Methods Biomech. Biomed. Eng. Imag. Visual. **6**(3), 283–292 (2018)
34. Xie, Y., Xing, F., Kong, X., Su, H., Yang, L.: Beyond classification: structured regression for robust cell detection using convolutional neural network. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 358–365. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_43
35. Zhang, L., Lu, L., Nogues, I., Summers, R.M., Liu, S., Yao, J.: DeepPap: deep convolutional networks for cervical cell classification. IEEE J. Biomed. Health Inform. **21**(6), 1633–1643 (2017)
36. Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: a survey. arXiv:1905.05055 [cs], May 2019