Post- print of: K. Janczyk, T. Neumann and J. Rumiński, "The influence of image masks definition on segmentation results of histopathological images using convolutional neural network," 2019 12th International Conference on Human System Interaction (HSI), Richmond, VA, USA, 2019, pp. 47-53, doi: 10.1109/HSI47298.2019.8942600.

The influence of image masks definition on segmentation results of histopathological images using convolutional neural network

1st Kamil Janczyk
Medical Engineering Division
Ship Design and Research Centre
and
Dept. of Biomedical Engineering
Gdansk University of Technology
Gdansk, Poland
kamil.janczyk@cto.gda.pl

2nd Tomasz Neumann
Dept. of Biomedical Engineering
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland

3rd Jacek Rumiński *Member IEEE*
Dept. of Biomedical Engineering
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland

***Abstract*—In the era of collecting large amounts of tissue materials, assisting the work of histopathologists with various electronic and information IT tools is an undeniable fact. The traditional interaction between a human pathologist and the glass slide is changing to interaction between an AI pathologist with a whole slide images. One of the important tasks is the segmentation of objects (e.g. cells) in such images. In this study, we apply U-net and V-net convolutional neural network models to perform image segmentation. In particular, we analyze the role of the contour thickness in the reference (labels, masks) images on the results of image segmentation, also for the degraded images. We show the role of the proper mask definition and the results obtained for the ensemble models that use the same architecture but are trained using two sets of inverted masks.**

***Index Terms*—histopathological images, convolutional neural network, annotation, edge detection, color-to-grayscale**

I. Introduction

The histopathological images are traditionally visually investigated by a trained pathologist using a microscope. The human pathologist interacts with a microscope to change the field of view, region-of-interest, etc. Recently, the introduction of professional scanners has enabled the reliable digitalization of the entire slide producing the whole-slide image (WSI). The analysis of WSI can be supported by computer-aided methods to improve image quality, to segment cells or detect contours of cells, etc. The higher level procedures supported by a computer can be focused on the detection of the clinically preferred hot spots in the WSI or to perform cells counting, etc. The artificial intelligence (AI) pathologist could perform the entire analysis of the WSI (detection, classification/recognition) producing the final description (WSI captioning using AI-based natural language processing). The traditional interaction between a human pathologist and a glass slide is changing to interaction between AI pathologist with a WHI. Digital whole-slide images are typically very large (GBs per image) so the automated, computer-based analysis is potentially very attractive. Therefore, many studies have been focused on processing of digital slide images.

A review of computer-assisted diagnosis technology for digitized histopathology was presented by Gurcan et al. in [1]. The paper describes issues related to histopatological images preprocessing (e.g. color and illumination normalization), image classification and segmentation. Another related survey focused on comparison of various colon cancer detection techniques categorized on the basis of adopted methodology and underlying dataset was presente dby Rathore in [2]. Most of the presented techniques have been evaluated on similar datasets showing that there are still many problems in image segmentation that should be addressed. A comparison of various segmentation techniques used for histopatological images was presented by Haj-Hassan et al [3]. Four techniques were compared: thresholding, edge-based segmentation, region-based segmentation and the snake-based (active-contours) method. The last method produced best results in detecting irregular shape as carcinoma cell type, achieving the average Dice and Jaccard similarity coefficient equal to 0.83 and 0.76 accordingly. Detecting cells nuclei using edge detection and multi-curvature cell nucleus contour model was proposed by Pang [4]. His solution was tested on the dataset containing 58 H&E stained images of breast cancer and achieved better results than other cell nuclei detection algorithms (for an exemplary image he achieved 10 correct detections using the proposed solution and only 5 correct detections with comparative methods). Albayrak and Bilgin proposed the application of the superpixel method for the pre-segmentation phase and they used convolutional neural networks (CNNs) for the final classification [5]. The proposed solution was tested with 810 histopathological images representing ten kidney renal cell carcinomas. The overall accuracy of the method was observed to be 0.9876. Another application of CNNs for pixel-wise region segmentation in histopatological images was investigated by Su et al. [6]. Their proposition achieved higher

average recall and precision values than using method based on texton histograms with logistic boosting and the application of the support vector machine method as texture classifier. Ma investigated semantic segmentation of histopatological images using U-net and fully convolutional neural network (FCN) [7]. He used 802 images of H&E stained slides of colon biopsy to define the training and test datasets. The obtained results of binary classificiation showed that FCN achieved higher F1-score and accuracy values. Segmentation of colon histopathology images with combination of support vector machine (SVM) and convolutional neural network was proposed by Li el at [8]. They used a subset of a dataset available from the Warwick Department of Computer Science. This dataset consists of 85 annotated images of H&E stained slides of glands [9], [10]. Their results showed that fusing GoogLeNet, AlexNet and hand-crafted features together with the SVM cassifier allowed to improve final segmentation results. The obtained values of Jaccard and Dice index were equal to 0.77 and 0.87 accordingly. Application of U-net for biomedical images segmentation with a few training images and emphasis on data augmentation was proposed by Ronnenberger et al. [14]. The proposed solution was tested on two datasets with partially annotated images. An average intersection over union (IoU) of 92% was achieved for 35 images of Glioblastoma-astrocytoma U373 cells recorded by phase contrast microscopy. The second dataset consisted of 20 images of HeLa cells, recorded by differential interference contrast microscopy. The obtained results for these images were characterized by an average IoU of 77.5%. So far, many different studies have been proposed to improve the segmentation method during the analysis of microscopic images. However, the role of reference image masks specification was not often addressed. In this study we focused on: a) the analysis of the role of the contour thickness in the reference (labels, masks) images on the results of image segmentation using CNN-based U-Net and V-Net models, b) the analysis of the negative vs. positive (binary) masks definition on image segmentation results, c) the influence of merging the segmentation results using two models trained with negative and positive masks on final segmentation accuracy, and d) the analysis of image degradation on segmentation results using pre-trained models. The rest of the paper is structured as follows. In Section II we describe the methodology used in our study. Results are presented in Section III and are discussed in Section IV. Conclusions are presented in Section V.

II. Method

Machine learning has now become a tool that quickly analyzes large sets of medical images, achieving comparable efficacy (sometimes even greater) than a specialized histopathologist. However, to recognize pathological states well, the CNN model needs a large collection of images in the learning phase with correct markings of hot spots, cells and other characteristic regions (e.g. pathologies). Therefore, in this work we would like to present the analysis of image segmentation as an important step allowing the description of important image regions. As a result a AI system can interact with a virtual slide to deliver important pre-processing results that could be further investigated by a trained histopathologist or a dedicated AI process.

A. Datasets

The database for our research consisted of 131 H&E histological images of colorectal adenocarcinomas, each 775x522px large, together with annotation images (masks). It is a part of the collection made available by the Warwick Department of Computer Science [9], [10]. The images were divided into two sets - a training set with 79 images and a test set containing remaining 52 images.

B. Data preparation

1) Training datasets: In the first step, H&E color images were converted to grayscale using the standard Matlab function *rgb2gray*, which uses specific weights for every channel in RGB image ($0.2989 \cdot R + 0.5870 \cdot G + 0.1140 \cdot B$). Then the images were cropped out into four 512x512px large images. This is the size of an image typically used for CNN-based segmentation algorithms. As a result of this pre-processing step the training dataset was increased from 79 to 316 (Fig. 1). The original "label" images were processed to investigate

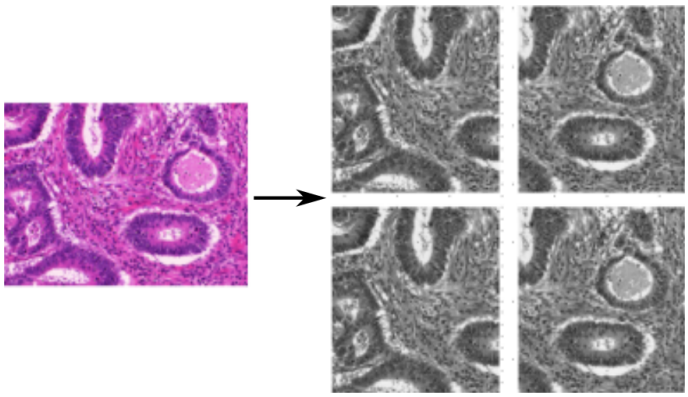

Fig. 1. The original image and the results of cropping in grayscale

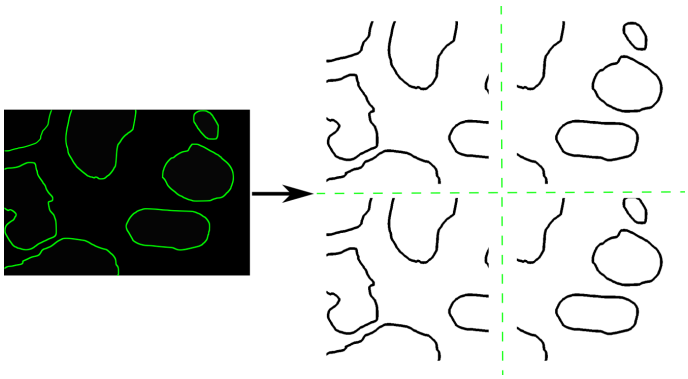

Fig. 2. The original "label" image with green lines representing borders of segments and generated four label images with the contour thickness of 7px

the role of masks definitions on final segmentation results. Two

main datasets of binary annotations were prepared - positive and negative (e.g. black contours on white background vs. white contours on black background). Each dataset consisted of the same number of images. The original masks were processed using binarization, followed by the Canny detector generating contours of the segmented objects. Afterwards the annotations images were also cut to 512x512px. (Fig.2). We generated five versions of positive and negative datasets, specifying different contour widths: 1px, 7px, 20px, 30px, and filled segments. The contour width was modified towards the center of the object being marked.

2) Test datasets: Base testing dataset was generated with the same Matlab function *rgb2gray* which was used for the training dataset. To test ability of neural network to segment noised images, different transformations were applied to the base test images. Blurred test dataset was generated using Matlab function *fspecial* and *imfilter*. The test datasets with applied Gaussian white noise were prepared using Matlab function *imnoise* - three different values of variance we used: 0.1, 0.05 and 0.025. The test dataset with impaired white balance was generated by reversing gray world algorithm (Fig. 3) [15]. Additionally, the different methods of conversion of a color image to a grayscale image were evaluated. First, the minimum and maximum decomposition algorithm was used, which correspondingly selects minimum and maximum values from RGB channels. Second, the mean algorithm was used, which depends of average values of RGB channels ($(R+G+B)/3$). Finally, a method that generates grayscale images using the lightness component from the CIE L*a*b* color space was used. Each test dataset originally consisted of 52 samples, but after the same division that was made to the training set, number of samples increased to 208. The ground truth images were prepared the same way as for the training dataset.

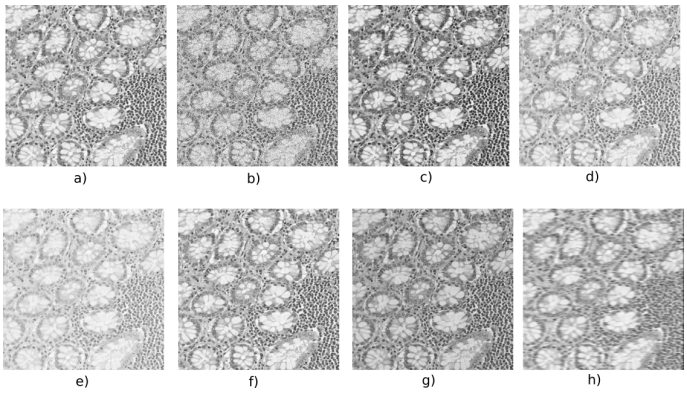

Fig. 3. Examples of test images: a) an original image, b) applied Gaussian noise with variance of 0.1, c) the image generated using minimal RGB values, d) the image generated using mean RGB values e) the image generated using maximal RGB values f) the image generated using lightness values g) the image with impaired white balance, h) the blurred image

C. The CNN models and training procedure

Three models of CNN [11], [12] were used in our study. The first model was based on U-net [13], which architecture was inspired by publication [14]. This model (UNET1) consisted of 24 convolution layers, 4 max-pooling, up-scaling and concatenate layers along with 2 dropout layers. The number of filters in convolution layers started from 64 in the 1st layer to 1024 in 9th and 10th layers and then dropped to 1 filter in 24th layer. The depth of UNET1 model was equal to 4. The ReLU activation function was used in most layers, except the last one were sigmoidal function was used. The Adam algorithm was used as an optimizer with learning rate $lr = 1e-4$. The loss function was defined as binary cross entropy. The second model was another implementation of U-net (UNET2). It consisted of 23 convolution layers for which the number of filters increased from 64 filters in first two layers to of maximum 1024 filters in 9th and 10th layers and then decreased to 1 filter in the last convolutional layer. After every convolutional layer there was batch normalization layer with exception of 11th, 14th, 17th and 20th convolution layers, after which there were concatenate layers. Model contained also 4 max-pooling and up-sampling layers as well as 1 dropout layer. The learning rate was equal to 0.1. Other training parameters were the same as for the first U-net model. Each model was described by about 31 millions of trainable parameters. The last model used in experiments was an application of two dimensional V-net - the fully convolutional neural network, developed by [16]. Similarly to the second model it consists of 23 convolution layers, 4 max-pooling layers and 4 up-scaling layers. The number of filters in the convolution layers increased from 6 to 256, and then decreased to 1. Model was divided into five stages, PReLu was used as an activation function in the first four stages. In the last stage activation took place using the sigmoidal function. This model used the Nadam algorithm with learning rate $lr = 2e-4$. The V-net model was described by about 24 million of trainable parameters. It is also important to underline, that for the UNET2 and V-net models we used the loss function defined based on the Dice coefficient. All three models were trained using two configurations. In the first configuration, the number of epoch was set to 120 and each epoch consisted of 300 steps with a batch size equals to 16. In the second settings, there were 60 epochs, and 2000 steps with a batch size of 2. For each configuration two annotations sets with reversed colors (named positive and negative) were evaluated. In total, 60 experiments were performed: 3 (models) * 5 (masks types) * 2 (positive vs. negative) * 2 (training configurations). All experiments were performed on the NVIDIA DGX Station platform using Keras (v. 2.2) with TensorFlow (v. 1.11) backend. The models trained for positive and negative versions of masks were used to generate the predicted segments for test images. For a given test image and for each configuration of experiments two images with predicted segments were obtained: segments for a model trained with "positive" masks and segments for a model trained with "negative" masks. The colors of an image with "negative" masks were inverted and merged with an image with positive masks. As a result an image with "merged" segments was produced. We analyzed if an image with "merged" segments represents better or similar

segmentation results in reference to segments produces using a model trained either with ”positive” or ”negative” masks. Finally, the trained models were tested using the modified (degraded) versions of test images to analyze how the models are sensitive to different changes in images.

III. Results

To analyze and evaluate data generated by convolutional neural network we computed the accuracy, recall, specificity, and F1-score parameters using a generated confusion matrix [17], [18]. Additionally, we used DICE and IoU metrics to compare the results. For all models, results obtained for the label images with a contour size of 1px were very poor quality (specificity or recall equal to 0 for a test dataset). Results obtained for other types of label images depend on the model type and the particular type of the label used in training. However, some models (16 out of 60) obtained for both U-net architectures were very poor (i.e., sensitivity or specificity was < 0.3). In particular, for UNET1 poor models were obtained for the following configurations: 1) for the negative set: a) for contour thickness 7px, b) for fully filled contours / masks (both settings configurations), 2) for the positive set: a) for contour thickness 7px, b) for contour thickness 20px, c) for masks (both configurations); for UNET2, for all models obtained using the second configuration (60 epochs, 2000 series, 2 images in an epoch) were poor with exception of label thickness 30 for ”negative” set. Additionally, poor models were generated using the positive set and settings configuration 1 (120 epochs, 300 series, 16 images in batch): a) for contour thickness 7px, b) 20px. The most stable model was V-net with acceptable results for all experiments.

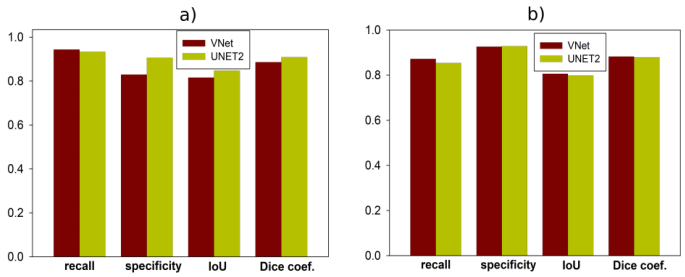

Fig. 4. Comparison of some quality metrics obtained for the V-net and U-net (UNET2) models for full (filled) segments using the positive (a) and negative (b) versions of masks.

The results obtained for the V-net model using two different training settings (i.e., configuration 1: 120 epochs, 300 series, 16 images in batch vs. configuration 2: 60 epochs, 2000 series, 2 images in an epoch) shown that differences in results were relatively small (up to 3.5%). Because of that we decided to further describe only the V-net model trained using parameters from the first configuration. We also used this model to investigate the behaviour of image segmentation performed on modified (degraded) test images.

The average values of recall (sensitivity) and specificity were calculated for segmentation results obtained using each test dataset with modified (degraded) images. The obtained results are presented in Table I, II (positive annotation set) and III, IV (negative annotation set). Values of standard deviation for obtained results were varied from 0.08 for masks (fully-filled contours) to 0.14 for contour size of 7px.

Due to the fact that we used corresponding ”negative” and ”positive” binary images with reversed colors, we could directly compare specificity calculated from models trained using ”positive” binary images with recall (sensitivity) calculated from models trained using ”negative” binary images.

Additionally, results for images with ”merged” segments (Fig.5) are presented in Table V and Table VI.

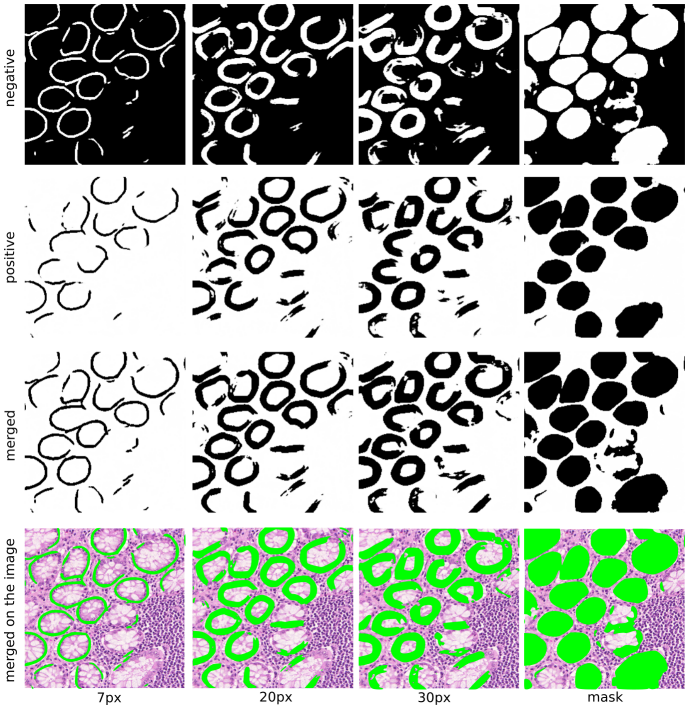

Fig. 5. Examples of images with segments obtained for: (from top) the model trained with ”negative” masks, the model trained with ”positive” masks, ”merged” segments, merged segments superimposed on the original image

TABLE I
Average values of specificity for different test sets using the ”positive” set

Added thickness	7px	20px	30px	mask
base images	0.43	0.65	0.63	0.83
Gaussian noise $\sigma^2 = 0.025$	0.43	0.64	0.62	0.84
Gaussian noise $\sigma^2 = 0.05$	0.41	0.62	0.61	0.84
Gaussian noise $\sigma^2 = 0.1$	0.36	0.55	0.57	0.83
blurred images	0.41	0.62	0.64	0.82
impaired white balance	0.34	0.53	0.56	0.82
maximum RGB values	0.22	0.37	0.40	0.66
minimum RGB values	0.35	0.56	0.54	0.83
mean RGB values	0.39	0.59	0.58	0.80
lightness values	0.44	0.65	0.63	0.83

TABLE II
AVERAGE VALUES OF RECALL FOR DIFFERENT TEST SETS USING THE "POSITIVE" SET

Added thickness	7px	20px	30px	mask
base images	0.98	0.96	0.97	0.94
Gaussian noise $\sigma^2 = 0.025$	0.98	0.96	0.97	0.94
Gaussian noise $\sigma^2 = 0.05$	0.98	0.96	0.97	0.93
Gaussian noise $\sigma^2 = 0.1$	0.98	0.96	0.97	0.91
blurred images	0.98	0.95	0.96	0.92
impaired white balance	0.98	0.96	0.96	0.87
maximum RGB values	0.99	0.97	0.98	0.96
minimum RGB values	0.98	0.96	0.97	0.89
mean RGB values	0.98	0.96	0.97	0.94
lightness values	0.98	0.96	0.97	0.95

TABLE III
AVERAGE VALUES OF SPECIFICITY FOR DIFFERENT TEST SETS USING THE "NEGATIVE" SET

Added thickness	7px	20px	30px	mask
base images	0.98	0.97	0.97	0.93
Gaussian noise $\sigma^2 = 0.025$	0.98	0.97	0.96	0.90
Gaussian noise $\sigma^2 = 0.05$	0.98	0.97	0.96	0.89
Gaussian noise $\sigma^2 = 0.1$	0.98	0.97	0.96	0.86
blurred images	0.98	0.97	0.95	0.91
impaired white balance	0.98	0.96	0.95	0.83
maximum RGB values	0.98	0.98	0.98	0.95
minimum RGB values	0.98	0.97	0.96	0.84
mean RGB values	0.98	0.97	0.97	0.91
lightness values	0.98	0.97	0.97	0.93

TABLE IV
AVERAGE VALUES OF RECALL FOR DIFFERENT TEST SETS USING THE "NEGATIVE" SET

Added thickness	7px	20px	30px	mask
base images	0.48	0.61	0.65	0.87
Gaussian noise $\sigma^2 = 0.025$	0.48	0.61	0.66	0.88
Gaussian noise $\sigma^2 = 0.05$	0.46	0.59	0.65	0.88
Gaussian noise $\sigma^2 = 0.1$	0.43	0.56	0.65	0.88
blurred images	0.45	0.61	0.65	0.85
impaired white balance	0.40	0.54	0.59	0.85
maximum RGB values	0.34	0.42	0.42	0.74
minimum RGB values	0.41	0.54	0.58	0.88
mean RGB values	0.46	0.58	0.60	0.85
lightness values	0.49	0.62	0.65	0.87

TABLE V
AVERAGE VALUES OF SPECIFICITY FOR DIFFERENT TEST SETS USING MERGED RESULTS

Added thickness	7px	20px	30px	masks
base images	0.55	0.72	0.72	0.90
Gaussian noise $\sigma^2 = 0.025$	0.55	0.72	0.73	0.90
Gaussian noise $\sigma^2 = 0.05$	0.53	0.70	0.72	0.90
Gaussian noise $\sigma^2 = 0.1$	0.49	0.65	0.69	0.90
blurred images	0.52	0.71	0.73	0.88
impaired white balance	0.46	0.63	0.66	0.88
maximum RGB values	0.38	0.49	0.50	0.77
minimum RGB values	0.47	0.65	0.65	0.90
mean RGB values	0.51	0.67	0.68	0.87
lightness values	0.55	0.72	0.72	0.89

TABLE VI
AVERAGE VALUES OF RECALL FOR DIFFERENT TEST SETS USING MERGED RESULTS

Added thickness	7px	20px	30px	masks
base images	0.97	0.95	0.95	0.91
Gaussian noise $\sigma^2 = 0.025$	0.97	0.95	0.95	0.89
Gaussian noise $\sigma^2 = 0.05$	0.97	0.95	0.95	0.87
Gaussian noise $\sigma^2 = 0.1$	0.97	0.95	0.95	0.83
blurred images	0.97	0.93	0.94	0.80
impaired white balance	0.97	0.94	0.93	0.89
maximum RGB values	0.98	0.96	0.97	0.94
minimum RGB values	0.97	0.95	0.94	0.82
mean RGB values	0.97	0.95	0.95	0.90
lightness values	0.97	0.95	0.95	0.92

A. Analysis of influence of contour thickness

Increasing the thickness of contours in reference images from 7px to 20px produced the significant improvement of recall and specificity in all cases - up to 0.22 difference in the average specificity for base images in "positive" set. Similar advance can be noticed for increasing thickness from 30px to fully-filled contour (masks). The thickness difference between 20px and 30px did not improve results vastly. Actually decrease in a few cases can be observed, mostly for images generated with the "positive" set.

B. Analysis of the influence of the added noise

Adding Gaussian noise to test images caused minor changes in the observed recall and specificity values. These changes were rising alongside with the increase of noise variance, with a few exceptions (Fig. 6). The above-mentioned exceptions include all results obtained for fully-filled contours (masks) and for contour thickness of 30px in the "negative" set. For these cases the average value of recall and specificity remains mostly the same (0.01 difference). The highest decrease for the average specificity value equals 0.1. It was caused by the Gaussian noise with 0.1 variance, for a contour thickness of 20px in the "positive" set. Image degradation with blurring produced similar effect as for the Gaussian noise. The minor decreases was observed in average values of recall and specificity for most cases. The high decrease of the average specificity (0.11) was observed for the contour thickness of 20px in merged results (Table III). Applying reversed gray world algorithm to test set caused higher (up to 0.12 for 20px contour thickness in the "positive" set) decrease of segmentation quality metrics. Mean decrease after impairing white balance, for all test sets, is equal to 0.07 ± 0.03 (Fig. 7).

C. Analysis of impact of different grayscale algorithms on contour detection

For most cases test images generated using the Matlab function *rgb2gray* and using the lightness value from the L*a*b* color space gave similar outcomes. The observed differences in average specificity and recall values were equal to 0.01. Results obtained using other grayscale conversion algorithms were worse, especially for maximum decomposition algorithm (up to 0.22 difference in the average specificity) (Fig. 8).

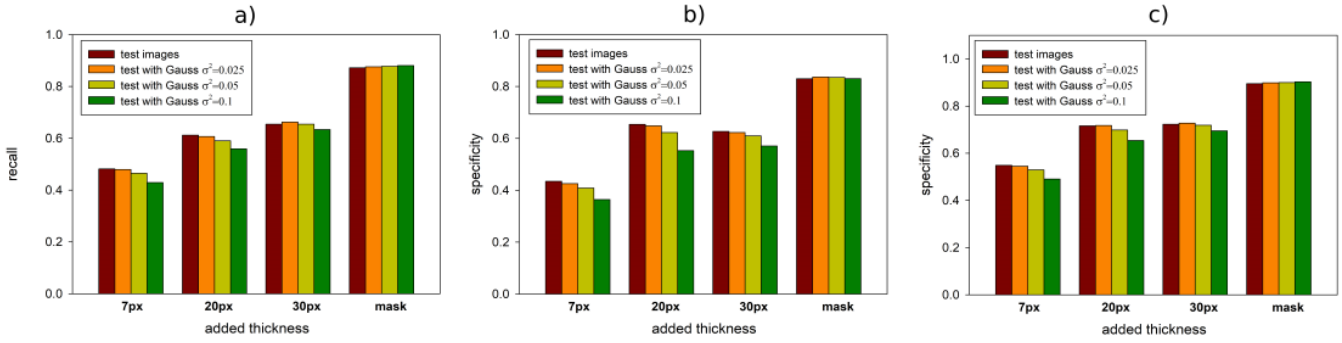

Fig. 6. Average values of recall and specificity for images degraded by a Gaussian noise: a) the ”negative” set, b) the ”positive” set, c) the merged segments

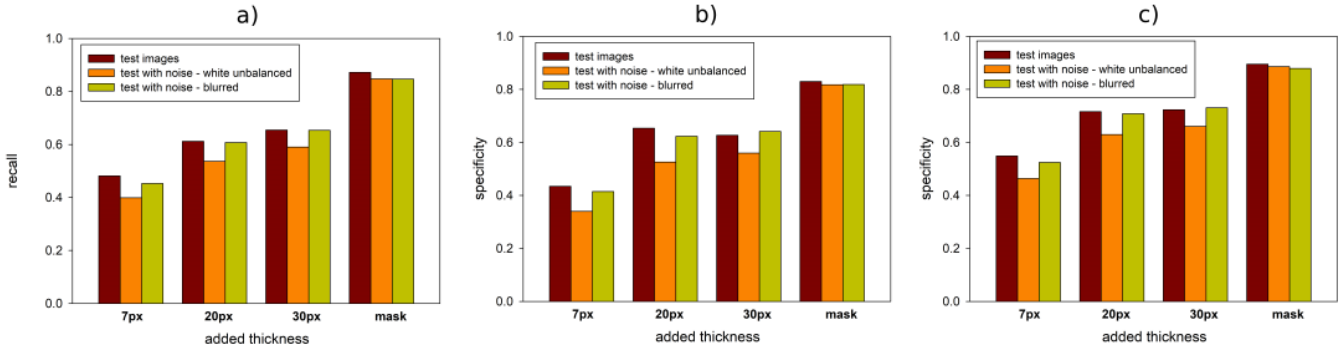

Fig. 7. Average values of recall and specificity for images degraded by impaired white balance and blurring: a) the ”negative” set, b) the ”positive” set, c) the merged segments

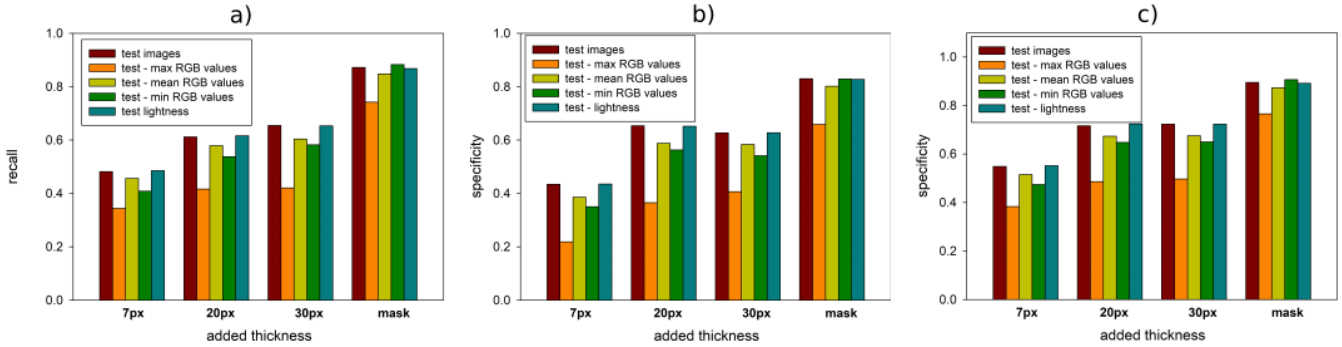

Fig. 8. Average values of recall and specificity obtained for images generated using different grayscale conversion algorithms: a) the ”negative” set, b) the ”positive” set, c) the merged segments

D. Analysis of impact of binary masks colors

Comparison of recall and specificity values obtained using different types of binary masks (”positive” vs. ”negative”) shows that training with the ”negative” masks (black background vs. white objects) allowed to achieve slightly better results. The best results were obtained for images with merged segments (up to 0.12 difference).

IV. Discussion

It was shown that the contour thickness used in mask (reference) images in training of U-net or V-net models plays a very important role. In general, if pixels of the label image (masks) are more equally distributed between classes (segments vs. background) then better segmentation results are obtained (recall/sensitivity vs. specificity). For examples, if contour thickness was 1px the trained models were unable to distinguish between 2 classes (all pixels were assigned to either class 0 or 1). Increasing contour thickness from 7px to 20px and from 30px to fully-filled masks led to higher values of recall and specificity. For a best case the average value of specificity increased by 0.48. The performed (repeated) experiments showed another important result. There

is a small, but observable difference between models trained using the "positive" and "negative" masks. Maybe, another model architecture should be proposed that allows obtaining identical results for such corresponding masks. In this work, we proposed to use two models trained separately using the "positive" and "negative" masks and to merge segments generated by two models. It was shown that images with merged segments improved the results for every test set. The highest specificity value for merged images was equal 0.90. Taking into account the quality and the content of analyzed images the obtained results are quite acceptable, but further improvement are still possible. The obtained results are comparable with other applications of convolutional neural network in segmentation of histopatological images [6], [7], including studies that used images from the same collection [8]. Another interesting observation from this study is the influence of image degradation on final segmentation results. In general, models trained for reference images with narrow contour thickness are more sensitive to image degradation. This is understandable since the percentage of pixels in such contours is relatively small in reference to background pixels so a small degradation in contour pixels lead to higher relative error. The models generated for masks and combined to produce the images with merged segments were not very sensitive to introduced image degradation. Adding Gaussian noise to the test images cause little deterioration of the results. For merged images, application of Gaussian noise did not influence the quality of results. The introduction of Gaussian noise with higher variance (e.g. 0.1) decreased the obtained values of quality metrics for models trained with label images that used narrow contours but did not change the results obtained for for models trained with full masks. Blurring produced only slightly worse results. Higher decrease of the segmentation quality metrics was obtained when images were degraded by color (gray) modifications. It was especially observed for the maximum decomposition procedure used to convert RGB images to grayscale images. The observed decrease of specificity for images with merged segments was about 0.13, which is much higher than for the second worst case (0.03).

V. Conclusions

It was shown that the contour thickness in the reference images (labels) plays an important role on the quality of segmentation using U-net/V-net models. Another interesting findings is a small, but observable difference between results obtained for models trained using the "positive" and "negative" masks. Theoretically, we could expect the corresponding results: the segments generated by the model trained with "positive" masks should be the inverted version of segments generated by the model trained with "negative" masks. The difference could be partially related to the fact that training is always a different task (e.g., randomly generated batches, initial model parameters, etc.). In this study, we shown that ensemble models (the same architecture but trained with inverse data labels) can produce better segmentation results. We also shown that limited image degradation due to the additive Gaussian noise or blurring practically did not decreased the segmentation accuracy obtained for ensemble models. Future research should be focused on expanding the database for both training and testing cases, e.g. use other methods of data augmentation and combining transformed images into one set. Other segmentation models (including instance segmentation) should be investigated especially to provide more balanced results for the "positive" and "negative" label datasets. Especially, U-net architectures need more precise analysis since there were very unstable, but in some experiments these models produced better results than for the V-net architecture.

References

[1] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener, "Histopathological Image Analysis: A Review", in IEEE Reviews in Biomedical Engineering, vol. 2, pp. 147-171, 2009.

[2] S. Rathore, M. Hussain, A. Ali and A. Khan, "A Recent Survey on Colon Cancer Detection Techniques", in IEEE/ACM transactions on computational biology and bioinformatics, vol. 10, pp. 545-563, 2013.

[3] H. Haj-Hassan, A. Chaddad, C. Tanougast and Y. Harkouss, "Comparison of segmentation techniques for histopathological images", in 2015 Fifth International Conference on Digital Information and Communication Technology and its Applications (DICTAP), pp. 80-85, 2015.

[4] B. Pang, L. Zhou, W. Zeng and X. You, "Cell Nuclei Detection in Histopathological Images by Using Multi-curvature Edge Cue", in 2011 Seventh International Conference on Computational Intelligence and Security, Hainan, pp. 1095-1099, 2011.

[5] A. Albayrak and G. Bilgin, "A Hybrid Method of Superpixel Segmentation Algorithm and Deep Learning Method in Histopathological Image Segmentation", Innovations in Intelligent Systems and Applications (INISTA), Thessaloniki, pp. 1-5, 2018.

[6] H. Su, F. Liu, Y. Xie, F. Xing, S. Meyyappan and L. Yang, "Region segmentation in histopathological breast cancer images using deep convolutional neural network", in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), New York, NY, pp. 55-58, 2015.

[7] Z. Ma, Z. Swiderska, N. Ing, H. Salemi, D. McGovern, B. Knudsen, and A. Gertych, "Semantic Segmentation of Colon Glands in Inflammatory Bowel Disease Biopsies", in Information Technology in Biomedicine, pp. 379-392, 2019.

[8] W. Li, S Manivannan, S. Akbar, J. Zhang, E. Trucco, S.J. McKenna, "Gland segmentation in colon histology images using hand-crafted features and convolutional neural networks", in IEEE 2016 IEEE 13th International Symposium on Biomedical Imaging, pp. 1405-1408, 2016.

[9] K. Sirinukunwattana, D. R. J. Snead and N. M. Rajpoot, "A Stochastic Polygons Model for Glandular Structures in Colon Histology Images," in IEEE Transactions on Medical Imaging, vol. 34, pp. 2366-2378, 2015.

[10] K. Sirinukunwattana, S.E.A. Raza, Y.W Tsang, I.A. Cree, D.R.J. Snead, N.M. Rajpoot, "Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images", in IEEE Transactions on Medical Imaging, vol. 35, pp. 1196-1206, 2016.

[11] M. Stuart and M. Manic, "Survey of progress in deep neural networks for resource-constrained applications", in Proc. 43rd Annual Conf. of the IEEE Industrial Electronics Society, IECON, pp. 7259-7266, 2017

[12] D. Marino, C. Wikramasinghe, M. Manic, "An Adversarial Approach for Explainable AI in Intrusion Detection Systems", in Proc. 44rd Annual Conf. of the IEEE Industrial Electronics Society, IECON, 2018

[13] zhixuhao, "Implementation of deep learning framework – Unet, using Keras", https://github.com/zhixuhao/unet, access date 2019-04-01.

[14] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", arXiv:1505.04597, 2015.

[15] G. Zapryanov, D. Ivanova, I. Nikolova, "Automatic White Balance Algorithms for Digital Still Cameras - a Comparative Study", in Information Technologies and Control, vol. 1, pp. 16-22, 2012.

[16] FENGShuanglang, "2D-Vnet-Keras", https://github.com/FENGShuanglang/2D-Vnet-Keras, access date 2019-04-01.

[17] Y. Sasaki, "The truth of the F-measure", in Teach Tutor Mater, 2007.

[18] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation", in Journal of Machine Learning Technologies, vol. 2, pp. 37-63, 2008.