A Dataset and a Methodology for Intraoperative Computer-Aided Diagnosis of a Metastatic Colon Cancer in a Liver

Dario Sitnik[a], Gorana Aralica[b,*], Mirko Hadžija[c], Marijana Popović Hadžija[c], Arijana Pačić[b], Marija Milković Periša[d,e], Luka Manojlović[b], Karolina Krstanac[b], Andrija Plavetić[e] and Ivica Kopriva[a,*]

[a]Division of Electronics , Ruđer Bošković Institute

Bijenička cesta 54, P.O. Box 180, 10002 Zagreb, Croatia

e-mail: Dario.Sitnik@irb.hr, ikopriva@irb.hr

[b]Department of Pathology and Cytology, Clinical Hospital Dubrava

Avenija Gojka Šuška 6, 10000 Zagreb, Croatia

e-mail: garalica@kbd.hr, arijanapacic@yahoo.com, lmanojlov@kbd.hr, karolina0101@windowslive.com

[c]Division of Molecular Medicine, Ruđer Bošković Institute

Bijenička cesta 54, P.O. Box 180, 10002 Zagreb, Croatia

e-mail: Marijana.Popovic.Hadzija@irb.hr, Mirko.Hadzija@irb.hr

[d]Clinical Hospital Center Zagreb, Kišpatićeva 12, 10000 Zagreb, Croatia

[e]Institute of Pathology, School of Medicine, University of Zagreb

Šalata 2, 10000 Zagreb, Croatia

e-mail: andrija.plavetic@gmail.com

Abstract

The lack of pixel-wise annotated images severely hinders the deep learning approach to computer-aided diagnosis in histopathology. This research creates a public database comprised of: (*i*) a dataset of 82 histopathological images of hematoxylin-eosin stained frozen sections acquired intraoperatively on 19 patients diagnosed with metastatic colon cancer in a liver; (*ii*) corresponding pixel-wise ground truth maps annotated by four pathologists, two residents in pathology, and one final-year student of medicine. The Fleiss' kappa equal to 0.74 indicates substantial inter-annotator agreement; (*iii*) two datasets with images stain-normalized relative to two target images; (*iv*) development of two conventional machine learning and three deep learning-based diagnostic models. The database is available at http://cocahis.irb.hr. For binary, cancer vs. non-cancer, pixel-wise diagnosis we develop: SVM, kNN, U-Net, U-Net++, and DeepLabv3 classifiers that combine results from original images and stain-normalized images, which can be interpreted as different views. On average, deep learning classifiers outperformed SVM and kNN classifiers on an independent test set 14% in terms of micro balanced accuracy, 15% in terms of the micro F_1 score, and 26% in terms of micro precision. As opposed to that, the difference in performance between deep classifiers is within 2%. We found an insignificant difference in performance between deep classifiers trained from scratch and corresponding classifiers pre-trained on non-domain image datasets. The best micro balanced accuracy estimated on the independent test set by the U-Net++ classifier equals 89.34%. Corresponding amounts of F_1 score and precision are, respectively, 83.67% and 81.11%.

Keywords: intraoperative diagnosis, metastatic colon cancer, liver, stain normalization, U-Net(++), DeepLabv3.

1. Introduction

According to the International Agency for Research on Cancer of the World Health Organization, there were 18 million new cases and 9.5 million cancer-related deaths in 2018 [1]. 10.2% of diagnosed cases are related to colorectal cancer. A recent publication predicts for the US population in 2020, 9% of colorectal cases among men and 8% among women [2]. The gold standard for diagnosing cancer is still microscopic examination through pathologist' visual inspection of stained histopathological samples [3, 4]. Due to the rise in cancer incidence, it is an increasingly complex and highly time-consuming task for pathologists [4-6]. That is why computer-aided diagnoses (CAD) are expected to relieve pathologists' workload [4-9].

Visibility of tissue structures is improved through staining histopathological samples with one or more dyes. The most frequently used dye by pathologists is hematoxylin-eosin (H&E). Fixation of tissue sample across a glass slide is carried out through formalin-fixed paraffin-embedding (FFPE). It typically takes 48 hours to prepare a glass slide for microscopic examination. Since the time for establishing diagnosis during surgery is minimal, rapid freezing of tissue followed by cutting on cryotome and shortly staining with HE is used instead of the FFPE section [8]. However, that influences tissue morphology and, consequently, the quality of staining, which, even in FFPE, is burdened with experimental variations known as batch effects [10]. That stands for motivation to develop a CAD-assisted intraoperative decision-making system. However, as pointed in [8], a small amount of research has analyzed frozen sections so far [11, 12]. The most important reason is the unavailability of publicly accessible datasets with a sufficient number of annotated histopathological images necessary to train classification algorithms. That is especially the case with pixel-wise annotated images. Because they require a large amount of annotated training data that creates a significant problem for convolutional neural networks (CNNs)-based deep learning (DL) structures [7, 5, 13-16].

In this paper, we introduce a database comprised of: (*i*) a dataset with 82 histopathology images of H&E stained frozen sections acquired intraoperatively on 19 patients diagnosed with metastatic colon cancer in a liver. Information is provided for each image whether it belongs to train or test set; (*ii*) corresponding pixel-wise ground truth maps annotated by four pathologists, two residents in pathology, and one student of medicine.; (*iii*) two datasets with images that were stain (color) normalized relative to two target images using a structure-preserving color normalization method [17]; (*iv*) a baseline multi-view like conventional machine learning and deep learning-based diagnostic models. Stain-normalized datasets are provided to cope with the problem caused by batch effects known as biochemical noise [19, 10]. The variations can change the quantitative morphological image features making it difficult to reach an accurate diagnosis for pathologists and CAD systems [10]. As emphasized in studies [20, 19], standardization of the H&E staining process is one of the critical prerequisites of computer-aided systems to produce accurate clinical data for use by anatomical pathology diagnosis assisting systems. A metastatic colon cancer histopathological annotation and diagnosis database is called CoCaHis. It is available at [21]. To the best of our knowledge, there is no other publicly available pixel-wise annotated dataset for this diagnosis.

In analogy with [6], we present in this paper the diagnostic performance of baseline pixel-wise CAD systems. They were designed to discriminate between cancerous and non-cancerous pixels and demonstrate the difficulty of the problem. Conventional machine learning-based CAD systems include incremental support vector machine (SVM) and adaptive k-nearest neighbors (kNN) classifiers. DL-based CAD systems include U-net [22], Nested U-Net (U-Net++), [23], with DenseNet201 as an encoder [24], and DeepLabv3 [25] classifiers. Motivated by the results presented in [9], to cope with the problem of insufficient training data, we initialized the U-net's and U-Net++' encoder weights (backbone) on the pre-trained ImageNet classification problem and then trained the whole network on the CoCaHis. Analogously, DeepLabv3 was pre-trained for segmentation problem on the PASCAL VOC 2012 dataset.

However, as shown in section 3, there was minor performance improvement relative to the case when weights of the corresponding networks were initialized randomly and trained from scratch. It is however, worth mentioning that all deep models converged faster in cases when pre-trained weights were initialized. On average, deep learning classifiers outperformed SVM and kNN classifiers on an independent test set 14% in terms of micro balanced accuracy, 15% in terms of the micro F_1 score, and 26% in terms of micro precision. As opposed to that, the difference in performance between deep classifiers is within 2%. The best micro balanced accuracy estimated on an independent test set, in the amount of 89.34%, was obtained by the U-Net++ classifier pre-trained on the ImageNet dataset. The corresponding amounts of F_1 score and precision are, respectively, 83.67% and 81.11%.

2. Materials and Methods

2.1 CoCaHis Database

The CoCaHis database contains 82 microscopic images of H&E stained sections of frozen human specimens of metastatic colon cancer in a liver collected intraoperatively from 19 patients. Thereby, 32.75% of the pixels represent cancer class. Images were collected through a clinical study from March 2017 to February 2020 at the Department of Pathology and Cytology in Clinical Hospital Dubrava, Zagreb, Croatia. The Institutional Review Board of the same hospital approved the collection of samples on May 24, 2016. All the patients gave written informed consent, and all the data were anonymized.

2.2 Staining

Pathologist established diagnosis of metastatic colon cancer in a liver after microscopic examination of samples stained by H&E and immunostained to diagnosis-relevant antigens using primary antibodies specific to: hepatocyte protein (Hep Par 1 - OCH1E5), a transcription factor expressed in colorectal carcinoma cells (CDX2), and cytokeratin 20 as a marker of adenocarcinoma cells (CK20 - clone Ks20.8, all from Dako, Denmark).

2.3 Image Acquisition

The image acquisition system was described previously in [25, 26]. To avoid overlap with the prior work, we present here a minimal quantity of technical details. The system is comprised of light microscope Olympus BX51 with a DP50 camera, UPPLANFL objective with 40× magnification and numerical aperture 0.75, and an eyepiece lens with a magnification of 10×. Thus, images were acquired with an overall magnification of 400×. The pathologist selected images with either mostly metastasis of colon cancer or with a combination of normal liver tissue and colon cancer metastasis focusing the camera manually by looking on the computer screen. The specimen was illuminated between 480 nm and 620 nm. RGB images with 2074 × 2776 pixels were acquired with an 8-bit resolution per monochromatic image. Afterward, images were down-sampled to 1388 × 1037 pixels. Since the pixel footprint' size is equal to 0.1098 μm^2 and the microscope spatial resolution is 0.45 μm, down-sampling did not cause the spatial information loss. However, it reduced the load on pathologists who performed pixel-wise labeling. Table 1 summarizes the distribution of images across the patients.

Table 1. Distributions of images per patient.

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
No. of images	15	8	8	5	6	4	4	2	2	1	4	2	8	4	1	1	2	2	5

2.4 Pixel-wise Labeling

Pixel-wise labeling of H&E stained sections images was done by four pathologists, two residents in pathology, and one final-year student of medicine. Labeling was performed on originally stained RGB images with the assistance of a super-pixels based software system that grouped similar pixels in a zoomed area. During the annotation process, annotators could choose between the brush and super-pixel tool. The size of brush or super-pixels and the super-pixel algorithm, e.g., SLIC, Watershed, Quickshift, and Felzenszwalb from the scikit-image package, [28], could be changed. During the annotation procedure, pathologists could fix incorrectly marked regions by discarding them or refining them. While the software system was reducing the burden on annotators by calculating super-pixels, they still could zoom the image and annotate it up to the pixel-level.

Pixel-wise labeling by seven experts was motivated by the known phenomenon of inter-observer variability and subjectivity [29]. Thus, ground truth labeling may suffer from the gold-standard paradox [30, 31]. Validation by multiple pathologists is required to minimize inter-observer variability and subjectivity [29]. To this end, we calculated the Fleiss' kappa statistics to estimate the inter-annotator agreement, i.e. reliability (validity) of annotation [32, 33]. To calculate the Fleiss' kappa statistics, we used freely available Matlab code [34]. The input to the function is a matrix $\mathbf{X}$ with the first dimension equal to the number of pixels, $1037 \times 1388 \times 82$, and the second dimension equal to 2, i.e. the number of classes. Thus, element x_{ij} stands for the number of annotators that at the pixel location i are declared to be cancerous (j=1) or non-cancerous (j=2). For a given dataset, the Fleiss' kappa statistic is equal to 0.74. Following the Table 2, adopted from [33], it corresponds to the substantial inter-annotator agreement. For each image separately, we also calculated the kappa statistics for seven annotators. That is shown in Fig. 1. The ground truth used for the experiments reported in section 3 was for each image obtained by the majority vote.

Table 2. Kappa statistics and strength of agreement.

Kappa statistics	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

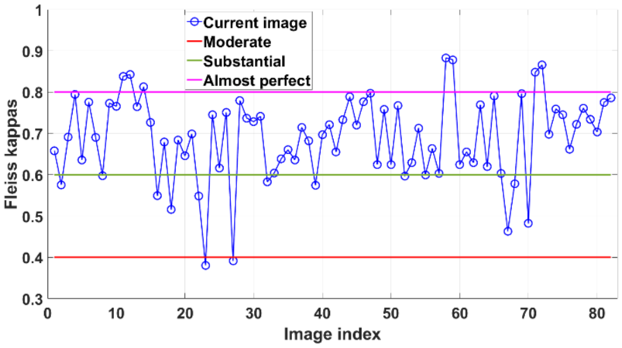


Fig. 1. The Fleiss' kappa statistics estimated for each image from pixel-wise annotations by seven experts.

As seen in Fig. 1, images 23 and 27 are characterized by a fair agreement between the annotators. Thus, we show in Fig. 2a the originally stained image 23, and in Fig. 2b to Fig. 2h the pixel-wise ground truth maps annotated by seven experts. Image 23 illustrates the challenges associated with intraoperative tissue collection, sectioning, and staining. In addition to the low quality of the frozen section, the cancer is poorly differentiated. Combining that with the color variation makes it hard to discriminate between the cancerous and non-cancerous pixels even for experienced pathologists. As it is already pointed out in [29], validation by multiple pathologists is required to minimize inter-observer variability and subjectivity. In such cases, the ground truth map used for the classifiers' training has to be formed by a majority vote of the greater number of annotators (seven in the study conducted herein). When the annotation is interpreted as a random process, the individual annotations can be interpreted as realizations of that process. Thus, the majority vote stands for the most probable outcome of the annotation process. Furthermore, shown in Fig. 1, images 11, 12, 14, 58, 59, 71 and 72 are characterized by an almost perfect agreement between the annotators. Therefore, we show in Fig. 3a the originally stained image 58, and in Fig. 3b to Fig. 3h, the pixel-wise ground truth maps annotated by seven experts. As opposed to image 23, cancer shown in image 58 is well-differentiated with clear boundaries corresponding to the fibrous tissue on the tumor's periphery. The quality of the frozen section is good. Since the experts' agreement is high, the majority vote does not play such an important role as in the previous case.

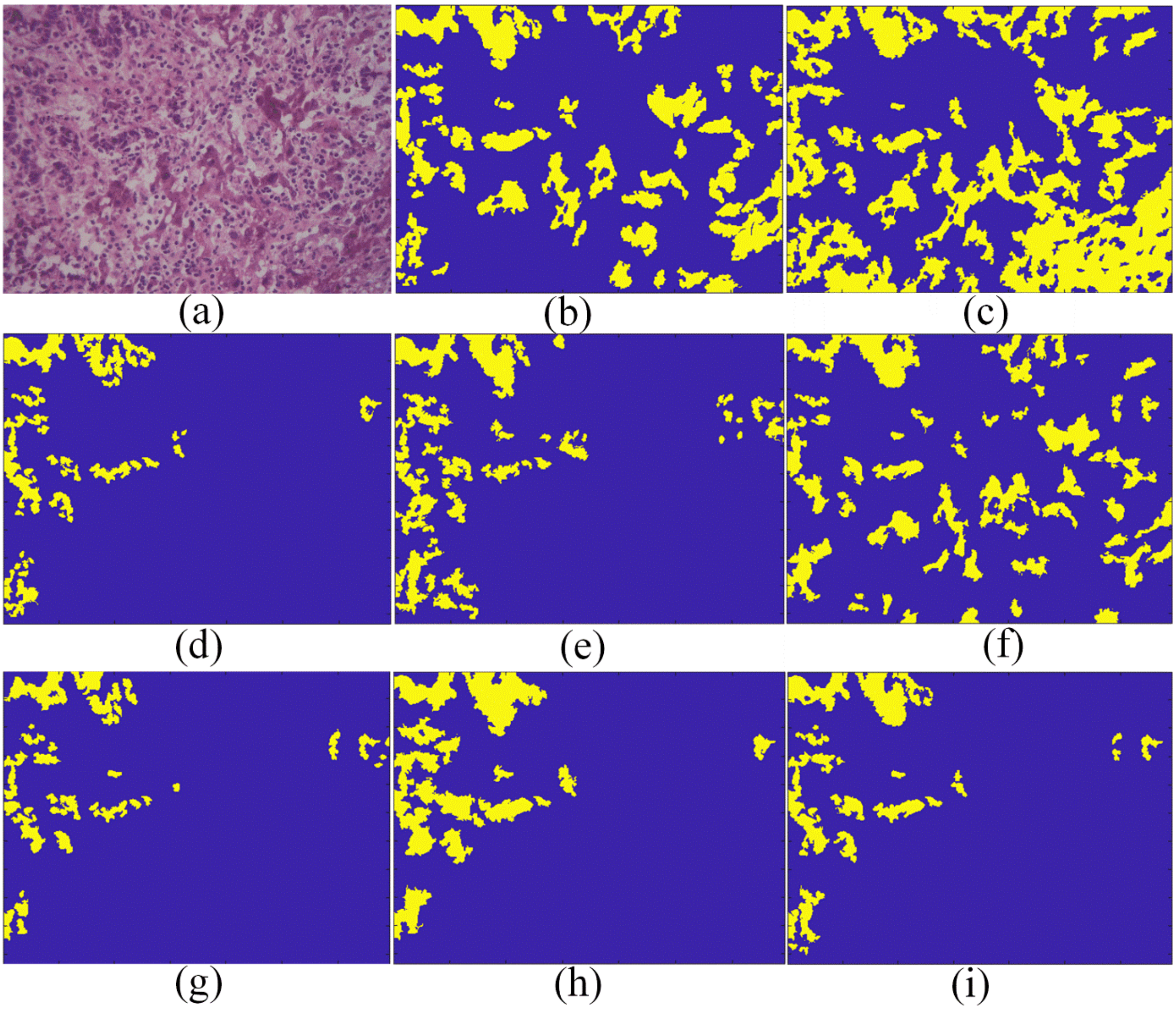


Fig. 2. (a) image of the H&E stained frozen section. (b) to (h) ground truth maps annotated pixel-wise by seven experts. Fleiss' kappa statistic equals 0.3805 and stands for a fair agreement. (i) majority vote ground truth map.

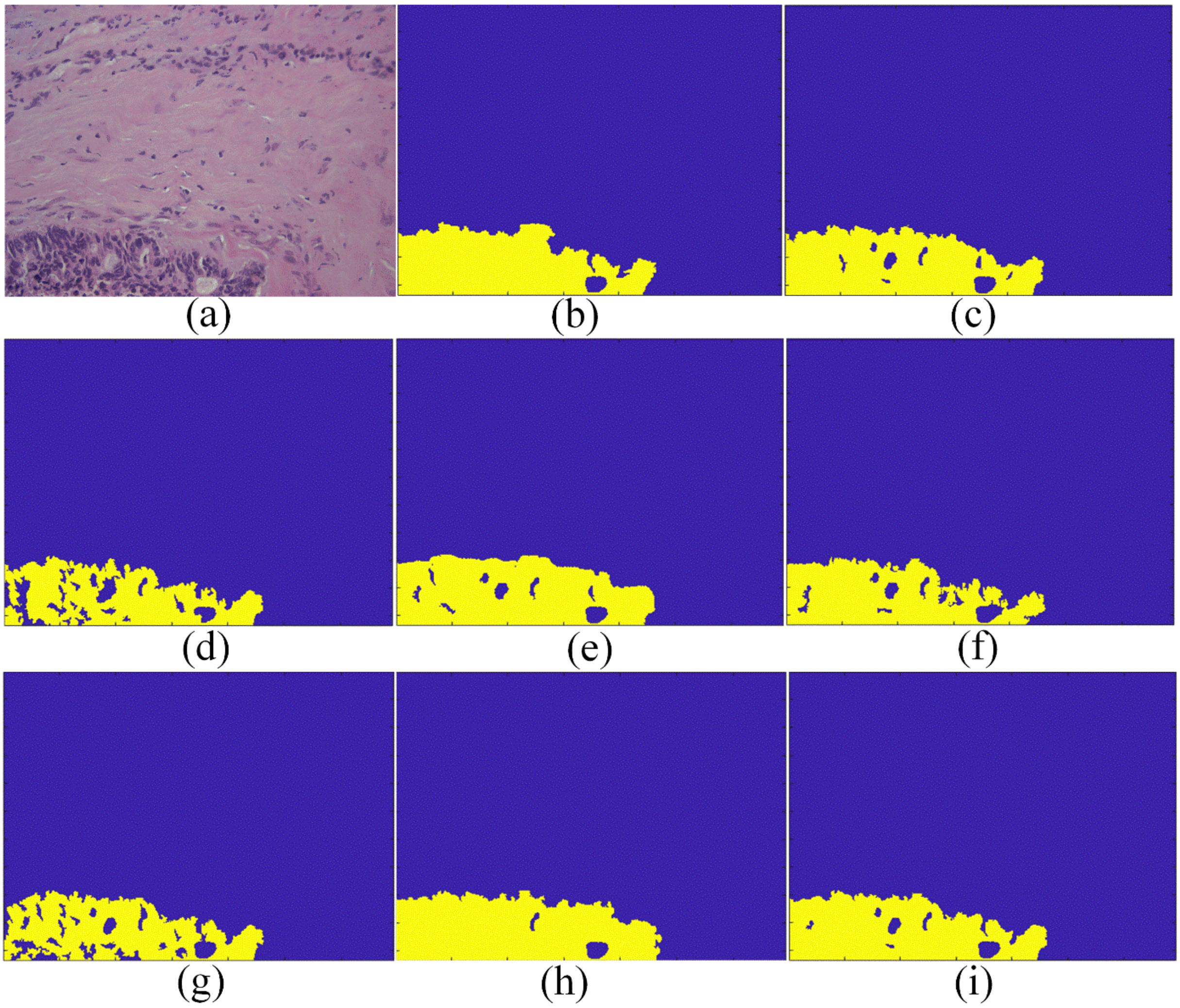


Fig. 3. (a) image of the H&E stained frozen section. (b) to (h) ground truth maps annotated pixel-wise by seven experts. Fleiss' kappa statistic equals 0.8829 and stands for almost perfect agreement. (i) majority vote ground truth map.

2.5 Stain Normalization

To cope with the experimental variations present during the slide preparation process, we applied the structure-preserved color (stain) normalization method [17], relative to target images selected by a pathologist, Figs. 4d and 4f. Thus, pixel-wise classification systems can be trained using either dataset composed of images of the H&E stained frozen sections or using one of the two datasets composed of stain-normalized images. As it was done in this paper, it is also possible to train classifiers on all three datasets and combine diagnostic results by majority vote. Fig. 4a shows one image of the H&E stained frozen section. The corresponding ground truth map is shown in Fig. 4b. Two images stained-normalized relative to target images Figs.4d and 4f are shown in Fig. 4c and Fig. 4e, respectively.

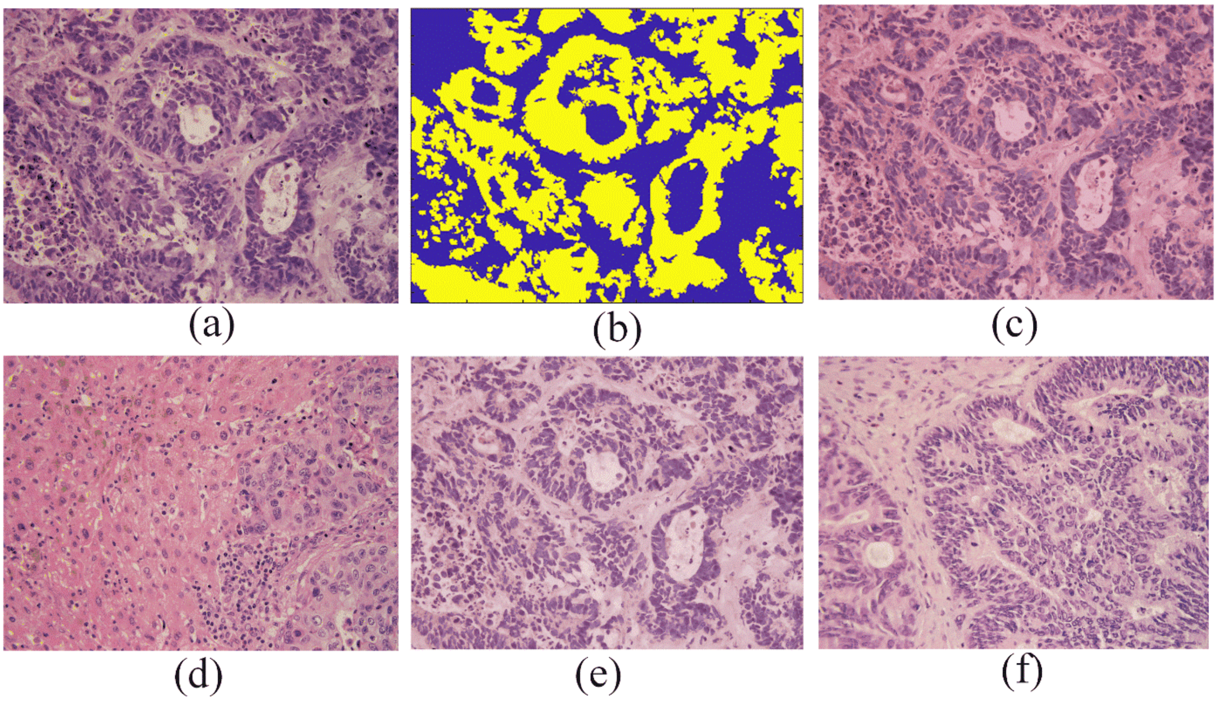


Fig. 4. (a) image of the H&E stained frozen section. (b) majority vote ground truth map. (c) stain-normalized image w.r.t. target image 1. (d) target image 1. (e) stain-normalized image w.r.t. target image 2. (f) target image 2. Images were acquired with magnification 400×.

2.6 Conventional Machine Learning Classifiers

Conventional machine learning classifiers such as SVM and kNN can yield good diagnostic performance if high-quality hand-crafted features are provided. It is however, known for a long time that "feature extraction matters more than the method used for classification" [35]. In other words, the extraction of discriminative features is a challenging problem in itself. Herein we relied on color as a feature that is a result of the H&E staining. In addition to the originally stained images, we trained the SVM and kNN classifiers on the two corresponding sets of the stain-normalized images. The biggest challenge in training these two classifiers in a pixel-wise setting comes from a huge training set (58 images with the size 1037×1388 pixels). Hence, we had to train the classifiers in an incremental (online) mode. Incremental learning of the linear SVM is implemented through the `incrementalLearner` function using Matlab 2020b. The learner can be implemented by one of the three types of solvers: scale-invariant solver [36], stochastic gradient descent (SGD) solver [37], and average SGD (ASGD) solver [38]. Linear SVM classifier is the only one supported by the `incrementalLearner` function. In our previous work [26] SVM classifiers were trained on the selected region of interest with the size of 100×100 pixels. The linear SVM classifier exhibited the best performance. That provides an additional justification for its use in the incremental mode herein. kNN is applied in an adaptive mode employing a sliding window algorithm (ADWIN) [39], using a scikit-multiflow package [40].

2.7 Deep Learning Classifier

Deep learning (DL) is a methodology that extracts feature representation directly from image data [5, 7, 15]. Because of that, DL is considered suited for image analysis challenges in digital pathology [41]. Herein, we are focused on fully convolutional neural networks (FCNNs). They differ from CNNs in that FCNNs replace the fully connected layer with the up-sampling and deconvolution layer [42, 14]. These layers are considered a backward version of, respectively, the pooling layer and convolutional layer. That yields a characteristic U-shape of the network [22]. FCNNs are adopted for image segmentation because they can be applied to images of virtually any sizes. Improvement of the well-known U-Net architecture for biomedical image segmentation [22], is known as U-Net++ architecture [23]. As shown in [9], pre-trained deep networks produced results comparable or even superior to results from state-of-the-art hand-crafted feature-based classification approaches. Thus, the insufficiency of labeled training data is addressed through (*i*) transferring the DensNet201 encoder network (backbone) [24] pre-trained on the *ImageNet* database for classification purposes. However, in section 3, we compared results achieved by pre-trained backbones of U-Net and U-Net++ networks with results achieved by networks trained from scratch. We found an insignificant difference in performance. See Tables 5 to 8, but the convergence speed improved on average by 15% in the case of pre-trained initializations. It is also seen that in the problem considered herein, the U-Net++ brought the minor performance improvement compared to the U-Net. (*ii*) augmenting data every epoch by elastic transformations, e.g., zoom, shear, rotation, horizontal and vertical flip, by randomly choosing their parameters [43].

Repeated combinations of max pooling and down-sampling (convolution striding) leads to the reduced feature resolution. This problem is resolved by removing the last several max-pooling layers of deep CNN and up-sampling filters in the subsequent convolutional layers. Filter up-sampling step is carried out through the insertion of holes (zeros) between the nonzero coefficients of the kernel, which is known as *atrous convolution* (also known as dilated convolution) [44]. This step brings no extra computational burden but produces denser feature maps at multiple scales. The approach is known as DeepLab [45]. Its computationally improved version that eliminates the post-processing step is known as DeepLabv3 [25]. Its TensorFlow implementation, [46], was used to train the corresponding diagnostic model. To cope with the labeled data insufficiency, weights of DeepLabv3 were transferred from the Pascal VOC 2012 dataset and further trained with the rest of the network on CoCahis dataset. Analogously to the U-Net and U-Net++, we also trained the DeepLabv3 from scratch for comparison. Similarly, as it was the case with the U-Net and the U-Net++, we found an insignificant difference in performance between the DeepLabv3 network trained from scratch and pre-trained on the Pascal VOC 2012 dataset, see Tables 9 and 10. However, the pre-trained version converged 10% faster than the network trained from scratch. All deep learning-based diagnostic models were implemented in Tensorflow [47] and Keras [48].

2.8 Train and Test Protocol

The CoCaHis dataset has been divided into a training (70%) and a testing (30%) set. This splitting ratio was applied to images and patients. To ensure that the classifiers generalize to unseen data, we guarantee that the test set's images did not come from the train set's patients. In other words, in addition to images, patients included in the test set were not included in the train set. Regarding the U-Net, U-Net++, and DeepLabv3 classifiers, the training set is preprocessed by normalizing pixels to the range from 0 to 1. Furthermore, to extract the most out of the given set, patching is performed by a 128×128 sliding window with 64 pixels strides. By this approach, more cancer-context could be caught [26]. 20% of the preprocessed train set was assigned to the validation set to have an independent performance metric for reducing the learning rate or the early stopping. After the models are trained, the test set was preprocessed to fit the implementation-specific input size and range constraint. Also, the window was stridden by 32 pixels leading to 16 context-unique segmentations. A minimum number of context-unique pixel classifications as cancer, necessary to mark the corresponding pixel as cancer, was optimized on train images and applied to the test set [26]. That resulted in a diagnostic map for each test image.

Regarding the incremental linear SVM classifier, we used 10-fold cross-validation to tune the hyperparameters. The scale-invariant solver was validated on standardized and non-standardized samples. For SGD and ASGD solvers additional hyperparameters were the batch size validated from the set {2, 3, 5, 7, 10, 15, 20}, and the strength of ridge regularization term validated from the set {10^{-6}, 5×10^{-6}, 10^{-5}, 5×10^{-5}, 10^{-4}}. Regarding the kNN classifier, with the ADWIN change detector, cross-validation was performed on the same train and validation set as for deep learning models. It was used to select the hyperparameters: the number of neighbors, the maximum number of samples that can be stored in one leaf node, and the maximum size of the window storing the last viewed samples [40, 39]. The respective sets were {3, 5, 7}, {30, 40, 50}, and {1000, 2000, 3000}. On the originally stained images the cross-validated parameters in respective orders were: 5, 30, and 1000. On images stain-normalized relative to the target 1 the parameters were: 7, 40, and 1000. On image stain-normalized relative to the target 2 the parameters were: 7, 40, and 1000.

2.9 *Performance Measures*

Diagnostic performance was quantified using five metrics: sensitivity, specificity, positive predicted value (PPV), F_1, and balanced accuracy (BACC). Sensitivity, also known as recall and true positive rate (TPR), is defined as:

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

where *TP* denotes the number of true positives (correctly diagnosed cancerous pixels) and *FN* denotes the number of false negatives (incorrectly diagnosed cancerous pixels). Specificity, also known as selectivity and true negative rate (TNR), is defined as:

$$TNR = \frac{TN}{TN + FP}$$

where *TN* denotes the number of true negatives (correctly diagnosed non-cancerous pixels) and *FP* denotes the number of false positives (incorrectly diagnosed non-cancerous pixels). PPV, also known as precision, is defined as:

$$PPV = \frac{TP}{TP + FP}$$

F_1 score, also known as the Dice coefficient, is the harmonic mean of *PPV* and *TPR* defined as:

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Since 32.75% of the dataset's pixels were cancerous (dataset is imbalanced), we used BACC, defined as the arithmetic mean between TPR and TNR, instead of standard accuracy that is biased towards dominating class. For all five metrics, value 0 indicates the worse and value 1 the best performance.

3. Results

All the classifiers were trained independently for the set of originally stained images, and for each of the two sets of stain-normalized images. Three individual diagnoses were also combined using the majority vote. The micro diagnostic performance of the incremental SVM classifier is presented in Table 3. We selected the version of the incremental linear SVM classifier that yielded the highest value of BACC. That occurred with the scale-invariant solver [36] with standardized samples. The micro diagnostic performance of the kNN classifier is presented in Table 4. The micro diagnostic performance for the U-Net trained, respectively, from scratch and pre-trained on the ImageNet datasets is presented in Table 5 and Table 6. Corresponding results for the U-Net++ classifier are presented in Tables 7 and 8, and for DeepLabv3 classifier in Tables 9 and 10. DL-based classifiers outperformed the conventional machine learning classifiers: 14% in terms of micro balanced accuracy, 15% in terms of the micro F_1 score, and 26% in terms of micro precision. As opposed to that, the difference in performance between deep classifiers is within the margin of 2%. Thus, there were no significant performance improvements due to the pre-training on non-domain image datasets. As seen in Tables 3 and 4, conventional classifiers benefited from stain-normalization. In other words, since SVM and kNN were based on "hand-crafted" (color) features, color normalization was essential for performance improvement. That was achieved on datasets comprised of stain-normalized images. Unlike conventional classifiers, DL-based classifiers were able to learn new features from originally stained images and from stain-normalized versions. That is why in several cases the majority vote improved diagnostic performance further. See Tables 6 to 10. To visualize the quality of the diagnosis by SVM, kNN, U-Net, U-Net++, and DeepLabv3 classifiers, we provide the following illustrations. Fig. 5 shows the U-Net++ best diagnostic performance for the originally stained image and stain-normalized images. Fig. 6 shows the best result achieved for the same case by the SVN, kNN, U-Net++, U-Net, and DeepLabv3. Corresponding results for the worse diagnostic performance are shown in Figs. 7 and 8. Figs. 5 and 7 combined with Tables 6 to 10 demonstrate that majority vote combinations of diagnosis based on the originally stained and stain-normalized images improve performance in terms of TPR, TNR, PPV, F_1 score, and/or BACC. Furthermore, it is important to notice in Fig. 8 that the DeepLabv3 classifier yielded anatomically more meaningful and more accurate diagnosis than U-Net and U-Net++ in a clinically demanding scenario. We contribute that to the field of view expansion property of the DeepLabv3 network that, in comparison with the U-Net/U-Net++, leads to denser feature maps.

Table 3. Micro diagnostic performance of incremental linear SVM on standardized originally stained (OS) and stain-normalized (SN) images. Best values are in bold.

Metric	**OS images**	**SN target 1**	**SN target 2**	**Majority vote**
TPR	**0.9151**	0.8031	0.7379	0.7969
TNR	0.4588	**0.7248**	0.7656	0.7220
PPV	0.3931	0.5279	**0.5467**	0.5234
F_1 Score	0.4235	0.6830	**0.7035**	0.6698
BACC	0.6869	**0.7640**	0.7516	0.7594

Table 4. Micro diagnostic performance of incremental kNN classifier on non-standardized originally stained (OS) and stain-normalized (SN) images. Best values are in bold.

Metric	**OS images**	**SN target 1**	**SN target 2**	**Majority vote**
TPR	0.7830	0.7707	**0.8917**	0.8245
TNR	0.6111	**0.7606**	0.4853	0.6345
PPV	0.4355	**0.5523**	0.3990	0.4636
F_1 Score	0.5597	**0.6434**	0.5513	0.5935
BACC	0.6971	**0.7657**	0.6885	0.7295

Table 5. Micro diagnostic performance of U-Net classifier on originally stained (OS) and stain-normalized (SN) images. The network is initialized randomly and trained from scratch. Best values are in bold.

Metric	**OS images**	**SN target 1**	**SN target 2**	**Majority vote**
TPR	0.8408	**0.8974**	0.7611	0.8250
TNR	0.8851	0.8221	0.9209	**0.9241**
PPV	0.7371	**0.8103**	0.7867	0.8063
F_1 Score	0.7855	**0.8161**	0.7737	0.8155
BACC	0.8630	0.8598	0.8410	**0.8746**

Table 6. Micro diagnostic performance of U-Net classifier on originally stained (OS) and stain-normalized (SN) images. Backbone of the network is pre-trained on the ImageNet dataset and fine-tuned on CoCaHis dataset. Best values are in bold.

Metric	OS images	SN target 1	SN target 2	Majority vote
TPR	0.8492	0.8159	0.8051	**0.8534**
TNR	0.8742	0.9186	0.9217	**0.9237**
PPV	0.7212	0.7934	0.7976	**0.8108**
F_1 Score	0.7780	0.8045	0.8013	**0.8316**
BACC	0.8617	0.8673	0.8634	**0.8886**

Table 7. Micro diagnostic performance of U-Net++ classifier on originally stained (OS) and stain-normalized (SN) images. The network is initialized randomly and trained from scratch. Best values are in bold.

Metric	OS images	SN target 1	SN target 2	Majority vote
TPR	**0.8760**	0.8336	0.7995	0.8423
TNR	0.8942	0.9247	0.9199	**0.9258**
PPV	0.7500	0.8091	0.7928	**0.8131**
F_1 Score	0.7874	0.8211	0.7961	**0.8274**
BACC	**0.8851**	0.8792	0.8597	0.8841

Table 8. Micro diagnostic performance of U-Net++ classifier on originally stained (OS) and stain-normalized (SN) images. Backbone of the network is pre-trained on the ImageNet dataset and fine-tuned on CoCaHis. Best values are in bold.

Metric	OS images	SN target 1	SN target 2	Majority vote
TPR	**0.8886**	0.8033	0.8145	0.8639
TNR	0.8838	0.9231	0.9167	**0.9229**
PPV	0.7451	0.8001	0.7893	**0.8111**
F_1 Score	0.8097	0.8018	0.8017	**0.8367**
BACC	0.8862	0.8632	0.8656	**0.8934**

Table 9. Micro diagnostic performance of DeepLabv3 classifier on originally stained (OS) and stain-normalized (SN) images. The network is trained from scratch. Best values are in bold.

Metric	OS images	SN target 1	SN target 2	Majority vote
TPR	0.8239	0.8032	0.8028	**0.8247**
TNR	0.8980	**0.9317**	0.9228	0.9303
PPV	0.7559	0.8183	0.7993	**0.8193**
F_1 Score	0.7884	0.8107	0.8011	**0.8219**
BACC	0.8610	0.8675	0.8628	**0.8775**

Table 10. Micro diagnostic performance of DeepLabv3 classifier on originally stained (OS) and stain-normalized (SN) images. The network is pre-trained on the Pascal VOC 2012 dataset challenge and fine-tuned on CoCaHis. Best values are in bold.

Metric	**OS images**	**SN target 1**	**SN target 2**	**Majority vote**
TPR	**0.8824**	0.8155	0.7810	0.8457
TNR	0.8879	0.9123	0.9172	**0.9193**
PPV	0.7510	0.7809	0.7833	**0.8006**
F_1 Score	0.8114	0.7978	0.7822	**0.8225**
BACC	**0.8852**	0.8639	0.8491	0.8825

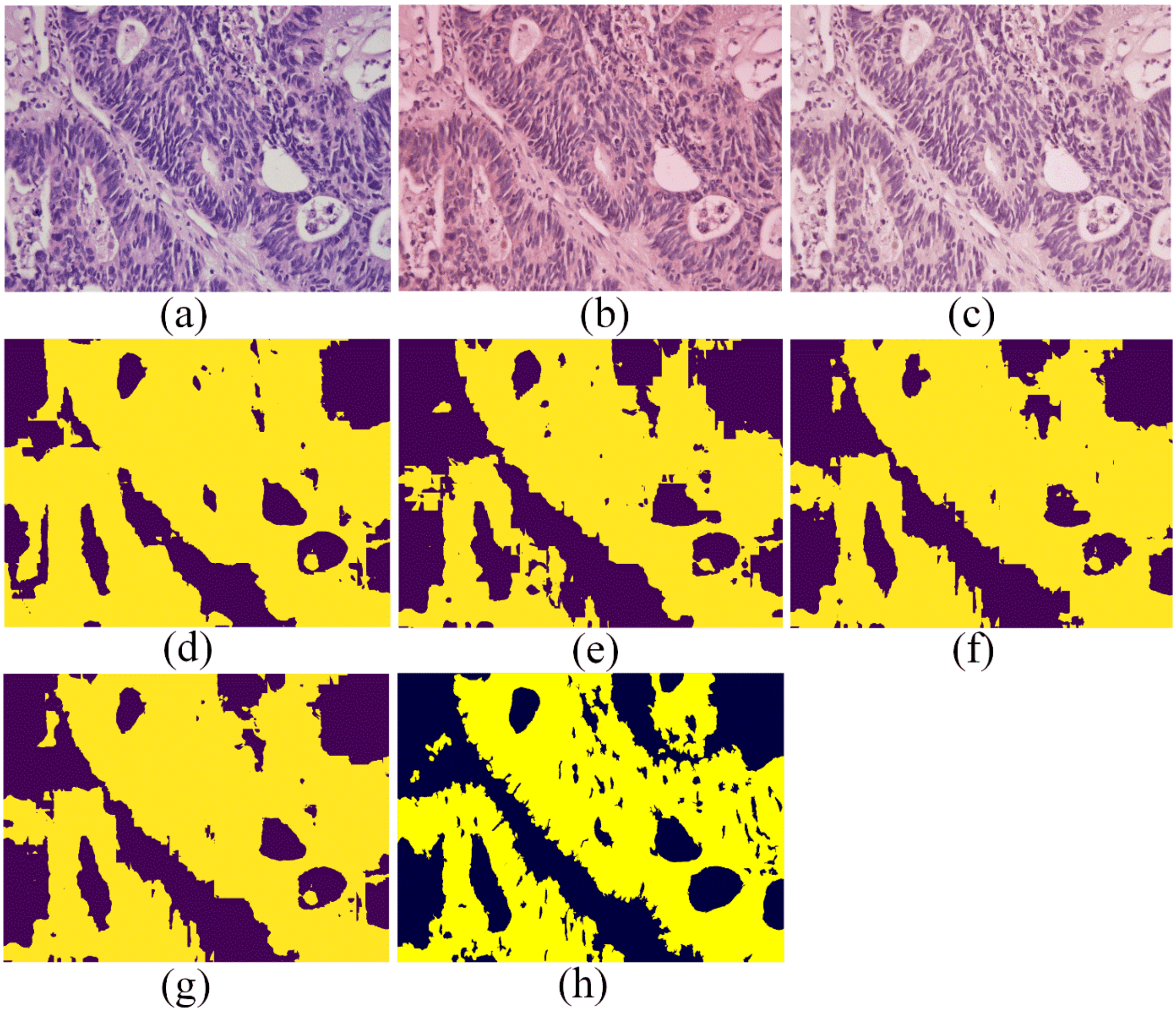


Fig. 5. (a) OS image 56 of H&E stained frozen section. (b) stain-normalized image w.r.t. target image 1. (c) stain-normalized image w.r.t. target image 2. (d) to (f) U-Net++ diagnosis. (g) majority vote diagnosis combining (d), (e) and (f). (h) majority vote ground truth. F_1 score between (g) and (h) equals to 0.9176. Images were acquired with magnification 400×.

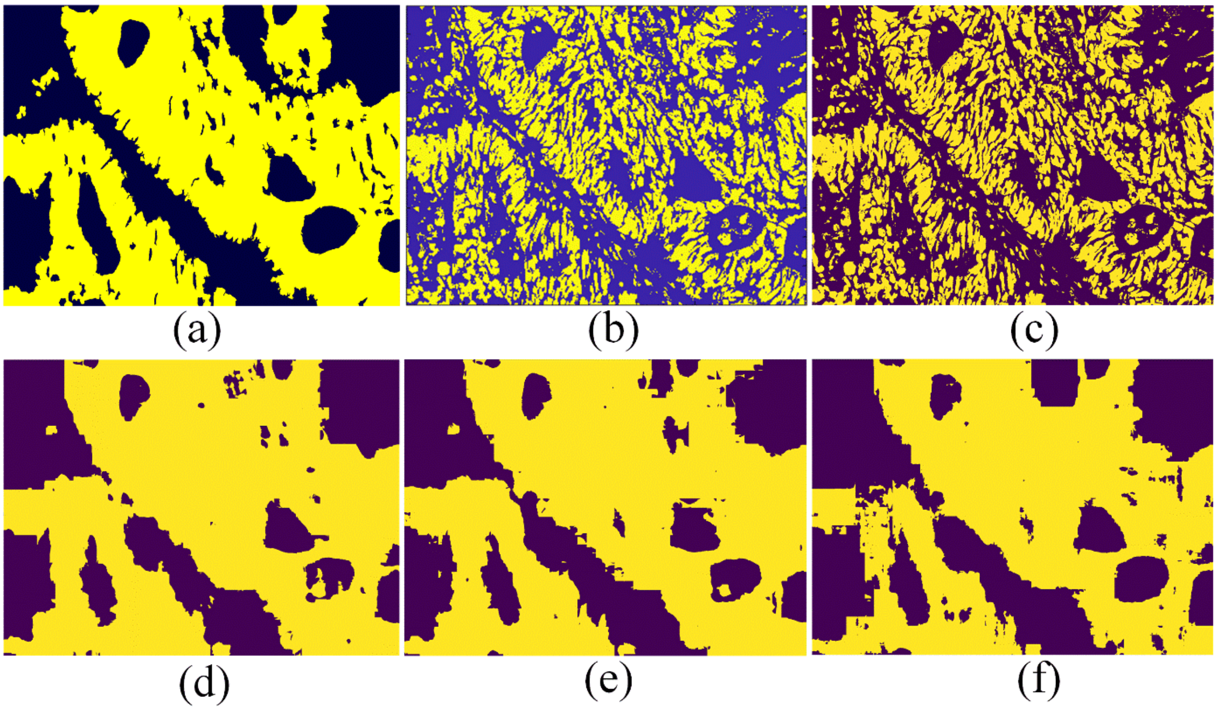
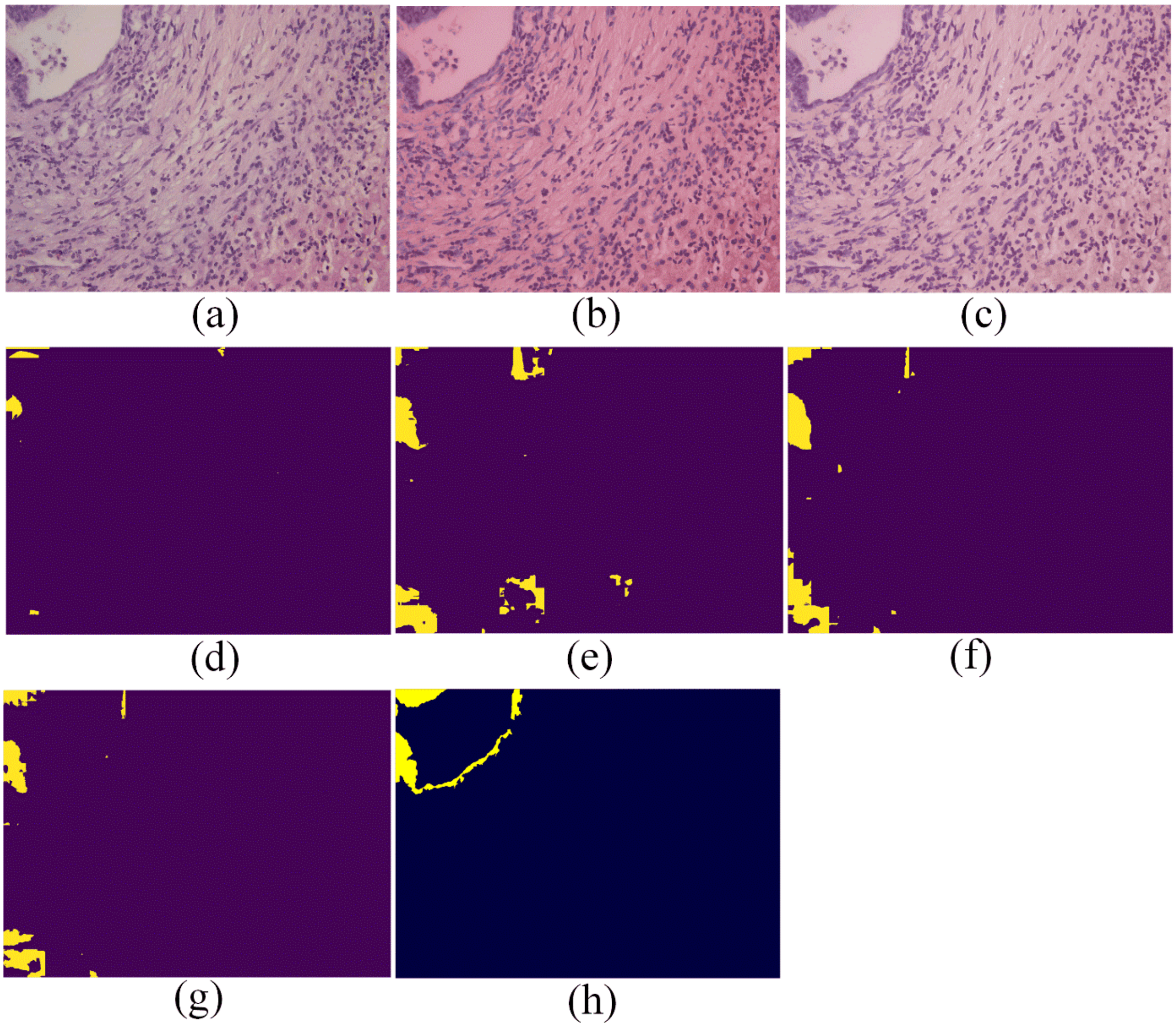


Fig. 6. Diagnostic maps for the images of H&E stained frozen section shown in Fig. 5(a), 5(b) and 5(c). (a) majority vote ground truth. (b) SVM' diagnostic map of stain-normalized image Fig. 5(b). (c) kNN diagnostic map on stain-normalized image Fig. 5(b). (d) Majority vote of U-Net++ diagnosis in Fig. 5 (d), 5(e) and 5(f). (e) U-Net, and (f) majority vote of DeepLabv3 diagnosis on original and stain-normalized images. F_1 scores of (a) and (b) to (f) in the respective order: 0.7202, 0.7219, 0.9176, 0.9171, and 0.9062.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 7. (a) image 45 of H&E stained frozen section. (b) stain-normalized image w.r.t. target image 1. (c) stain-normalized image w.r.t. target image 2. (d) to (f) U-Net++ diagnosis. (g) majority vote diagnosis combining (d), (e) and (f). (h) majority vote ground truth. F_1 score between (g) and (h) equals to 0.5141. Images were acquired with magnification 400×.

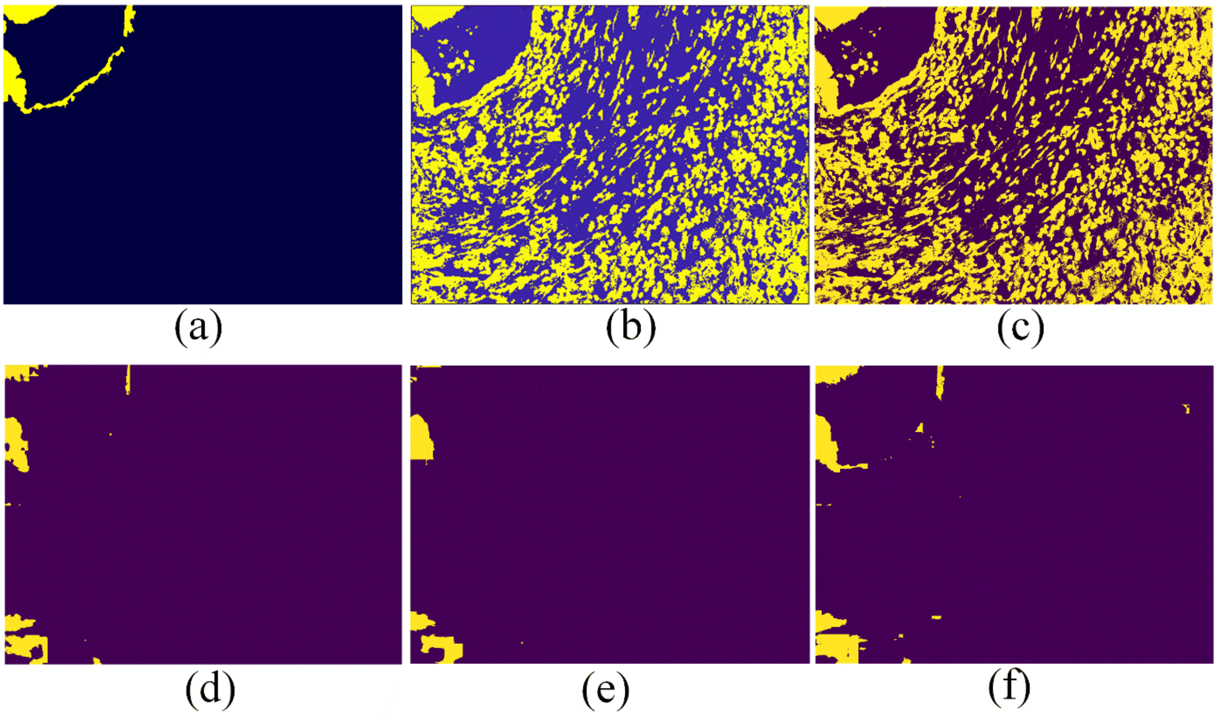


Fig. 8. Diagnostic maps for the images H&E stained frozen section shown in Fig. 7(a), 7(b) and 7(c). (a) majority vote ground truth. (b) SVM on stain-normalized image Fig. 7(b). (c) kNN on stain-normalized image Fig. 7(b). (d) U-Net++ by majority vote of Fig. 7(d), 7(e) and 7(f). (e) Majority vote of U-Net' diagnosis 7(d), 7(e) and 7(f). (f) DeepLabv3 by majority vote of diagnosis on original and stain-normalized images in Fig. 7. F_1 scores of (a) and (b) to (e) in the respective order: 0.0944, 0.0.0936, 0.5141, 0.3888, and 0.5674.

4. Discussion and conclusion

Artificial intelligence and computational pathology hold high potential in assisting pathologists in establishing diagnosing and/or grading cancer. However, that is hindered by the lack of publicly accessible datasets with expert labeling. Mostly, it is the case with an intraoperative pixel-wise diagnosis. Although CAD-assisted intraoperative decision-making systems are of potentially high clinical importance, their development is complicated due to several challenges: (*i*) collection of samples in the intraoperative procedure is demanding, (*ii*) the quality of frozen sections and quality of staining varies significantly due to tight time constrain imposed on laboratory technicians, (*iii*) highly time-consuming pixel-wise annotation process. Experimental variations in sample preparation, combined with the occasionally poorly differentiated cancer, make pixel-wise annotation even harder and more time-consuming, possibly leading to significant disagreement between the annotators. Thus, labeling by multiple pathologists is necessary to obtain a reasonably good estimate of the most probable outcome of the annotation process. That, however, adds further to the complexity of the development of the CAD-based intraoperative diagnosis systems. Motivated by outlined reasons, we provide a pixel-wise annotated database comprised of 82 histopathological images of H&E stained frozen sections. The sections with the metastatic colon cancer in a liver were collected intraoperatively on 19 patients. Seven experts generated corresponding pixel-wise ground truth maps. In addition to a dataset with originally stained images, two datasets comprised of images stain-normalized relative to two target images are also provided. Adding these two datasets is motivated by recent advances in computational methods for segmentation and/or classification of multi-view and/or multimodal data [49]. In principle, these methods improve performance relative to the case when either originally stained images or stain-normalized images are treated separately. Moreover, it is shown that SVM and kNN classifiers benefited from stain normalization. That is because they are based on "hand-crafted" (color) features. Due to the capability of learning new features, deep learning classifiers benefited from the stain normalization in a different way. They were able to learn discriminative and up to a certain extent, complementary features from all three datasets. That is why in several cases majority vote furtherly improved the diagnostic performance. Hence, a possible direction for further improving diagnostic performance is combining originally stained and stain-normalized versions into one multispectral image or one multi-view dataset. One of the challenges in the design of a CAD-based diagnostic system is the selection of the classifier. Herein, we verified that deep learning classifiers outperform SVM and kNN classifiers on an independent test set by a large margin in terms of balanced accuracy, F_1 score, and precision. It is also verified that the difference in performance between U-Net, U-Net++ and DeepLabv3 classifiers is within the margin of 2% independently on whether they were trained from scratch or pre-trained on non-domain image datasets. Thus, to clarify issues related to the architecture of deep learning networks and the usefulness of pre-training on non-domain image datasets, a larger database than the one presented herein is required. Currently, the baseline diagnostic result achieved on the independent test set is obtained with the "multi-view" U-Net++ classifier pre-trained on the ImageNet dataset. It respectively yields micro balanced accuracy, F_1 score, and precision in the amounts of 89.34%, 83.67% and 81.11%. Thus, the database can be used to train and test pixel-wise computer-aided intraoperative diagnostic systems that could act as a second reader.

CRediT authorship contribution statement

Dario Sitnik: Software for pixel-wise annotations, implementations of U-Net, U-Net++ and DeepLabv3 classifiers, implementation of adaptive kNN classifier, Writing the paper. **Gorana Aralica:** Study design, Sample collection, Pixel wise labeling. **Arijana Pačić:** Sample collection, Pixel wise labeling. **Marijana Popović Hadžija:** Sample collection. **Mirko Hadžija:** Sample collection, Image acquisition, Pixel wise labeling. **Marija Milković Periša:** Pixel wise labeling. **Luka Manojlović:** Pixel wise labeling. **Karolina Krstanac:** Pixel wise labeling. **Ivica Kopriva:** Funding acquisition, Study design, Study supervision, Implementation of stain normalization, MATLAB implementation of incremental SVM classifier, Writing the paper.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the QUADRO P6000 GPU used for this research.

Funding: This work was supported by the Croatian Science Foundation [grant number IP-2016-06-5235].

References

1. Cancer Today, IARC, 2018. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf

2. Siegel RL, Miller KD, Jernal A: Cancer Statistics. Cancer J Clin 2020, 70: 7-30. http://dx.doi.org/10.3322/caac.21590

3. Rubin's Pathology Clinicopathologic Foundations of Medicine, 6th Edition. Edited by Rubin R, Strayer DS, Philadelphia, PA, USA: Williams & Wilkins, 2012.

4. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B:Histopathological Image Analysis: A Review. IEEE Rev. Biomed. Eng. 2009, 2: 147-171. http://dx.doi.org/10.1109/RBME.2009.2034865

5. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, Hulsbergen-van de Kaa C, Bult P, van Ginneken B, van der Laak J: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports 2016, 6: e26286. http://dx.doi.org/10.1038/srep26286

6. Spanhol FA, Oliveira LS, Petitjean C, Heutte L: A Dataset for Breast Cancer Histopathological Image Classification. IEEE Trans. Biomed. Eng. 2016, 63: 1455-1462. http://dx.doi.org/10.1109/TBME.2015.2496264

7. Xing F, Xie Y, Su H, Liu F, Yang L: Deep Learning in Microscopy Image Analysis: A Survey. IEEE Trans. Neural Net. Learn. Syst. 2018, 29: 4550-4568. http://dx.doi.org/10.1109/TNNLS.2017.2766168

8. Komura D, Ishikawa S: Machine Learning Methods for Histopathological Image Analysis. Comp. and Struct. Biotech. J. 2018, 16: 34-42. http://dx.doi.org/10.1106/j.csbj.2018.01.001

9. Janowczyk A, Madabhushi A: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected cases. J. Path. Inform. 2016, 7: e29. http://dx.doi.org/10.4103/2153-3539.186902

10. Kothari S, Phan JH, Stokes TH, Osunkoya AO, Young AN, Wang MD: Removing batch effects from histopathological images for enhanced cancer diagnosis. IEEE J. Biomed. Health Inf. 2014, 18: 765-772. http://dx.doi.org/10.1109/jbhi.2013.2276766

11. Abas FH, Gokozan HN, Goksel B, Otero JJ, Gurcan MN: Intraoperative neuropathology of glioma recurrence: cell detection and classification. Proc. SPIE 9791, Med. Imag. 2016 - Digital Pathology, 9791: e979109. http://dx.doi.org/10.1117/12.2216448

12. Manni F, van der Sommen F, Zinger S, Kho E, Brouwer de Koning S, Ruers T, Shan C, Schleipen J, de With PHN: Automated tumor assessment of squamous cell carcinoma on tongue cancer patients with hyperspectral imaging. Proc. SPIE 10951, Med. Imag. 2019 - Image-Guided Proc., Robotics Interventions and Modeling, 10951: e109512K. http://dx.doi.org/10.1117/12.2512238

13. Xu X, Wu Q, Wang S, Liu J, Sun J, Cichocki A: Whole Brain fMRI Pattern Analysis Based on Tensor Neural Network. IEEE Access 2018, 6: 29297-29305. http://dx.doi.org/10.1109/ACESS.2018.2815770

14. Qiu JX, Yoon HJ, Fearn PA, Tourassi GD: Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports. IEEE J. Biomed. Health Inf. 2018, 22: 244-251. http://dx.doi.org/10.1109/JBHI.2017.2700722

15. Hu Z, Tang J, Wang Z, Zhang K, Zhang L, Sun Q: Deep Learning for Image-based Cancer Detection and Diagnosis - A Survey. Pattern Recognition 2018, 83: 134-149. http://dx.doi.org/10.1016/j.patcog.2018.05.014

16. Veta M, Pluim KP, van Diest PJ, Viergever MA: Breast cancer histopathology image analysis: A review. IEEE Trans. Biomed. Eng. 2014, 61: 1400-1411. http://dx.doi.org/10.1109/TBME.2014.2303852

17. Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, Steiger K, Schlitter AM, Esposito I, Navab N: Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. IEEE Trans. Med. Imag. 2016, 35: 1962-1971. http://dx.doi.org/10.1109/TMI.2016.2529665

18. Castleman KR, Eils R, Morrison L, Piper J, Saracoglu K, Schulze M, Speicher MR: Classification accuracy in multiple color fluorescence imaging microscopy. Cytometry 2000, 41: 139-147.
http://dx.doi.org/10.1002/1097-0320(20001001)41:2<139::AID-CYTO9>3.0.CO;2-N

19. Gavrilovic M, Azar JC, Lindblad J, Wählby C, Bengtsson E, Busch C, Carlbom IB: Blind color decomposition of histological images. IEEE Trans. Med. Imag. 2013, 32: 983-994. http://dx.doi.org/10.1109/TMI.2013.2239655

20. Bejnordi BE, Litjens G, Timofeeva N, Otte-Höller I, Homeyer A, Karssemeijer N, van der Laak JAWM: Stain Specific Standardization of Whole-Slide Histopathological Images. IEEE Trans. Med. Imag. 2016, 35: 404-415. http://dx.doi.org/10.1109/TMI.2015.2476509

21. CoCaHis datset. [online] at: http://cocahis.irb.hr

22. Ronneberger O, Fischer P, Brox T: U-net: Convolutional networks for biomedical image segmentation. Proc. Int. Conference on Medical Image Computing and Computer-assisted Intervention-MICCAI 2015, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28

23. Zhou Z, Mahfuzur M, Siddiquee R, Tajbakhsh N, Linag J: Unet++: A nested u-net architecture formedical image segmentation. International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DL MIA 2018, ML CDS 2018), LNCS 11045: 3-11. http://dx.doi.org/10.1007/978-3-030-00889-5_1

24. Huang G, Zhang L, van der Maaten L, Weinberger KQ: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 21-26, 2017, pp. 2261-2269. http://dx.doi.org/10.1109/CVPR.2017.243

25. Chen LI, Papandreou G, Schroff F, Adam H: Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv, 1706.05587v3, 2017.

26. Sitnik D, Kopriva I, Aralica G, Pačić A, Popović Hadžija M, Hadžija M: Transfer Learning Approach for Intraoperative Pixel-based Diagnosis of Colon Cancer Metastasis in a Liver from Hematohylin-Eosin Stained Specimens. Proc. SPIE 11320, Med. Imag. 2020 -Digital Pathology, 11320: e13200A. http://dx.doi.org/10.1117/12.2538303.

27. Kopriva I, Popović Hadžija M, Hadžija M, Aralica G: Unsupervised segmentation of low-contrast multichannel images: discrimination of tissue components in microscopic image of unstained specimen. Scientific Reports 2015, 5: 11576. http://dx.doi.org/10.1038/srep11576.

28. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, the scikit-image contributors: scikit-image: Image processing in Python. PeerJ 2014, 2:e453. http://dx.doi.org/10.7717/peerj.453

29. Colling R, Pitman H, Oien K, Rajpoot N, Macklin, CM-Path AI in Histopathology Working Group, Snead D, Sackville T, Verrill C: Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. J. Pathology 2019, 249: 143-150. http://dx.doi.org/10.1002/path.5310

30. Abels E, Pantanowitz L, Aeffner F, Zarella MD, van der Laak J, Bui MM, Vemuri VN, Parwani AV, Gibbs J, Agosto-Arroyo E, Beck AH, Kozlowski C: Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J. Pathol. 2019, 249: 286-294. http://dx.doi.org/10.1002/path.5331

31. Aeffner F, Wilson K, Martin NT, Black JC, Hendriks CLL, Bolon B, Rudmann DG, Gianani R, Koegler SR, Krueger J, Young GD: The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. Arch. Pathol. Lab. Med. 2017, 141: 1267-1275. http://dx.doi.org/10.5858/arpa.2016-0386-ra

32. Fleiss JL: Measuring nominal scale agreement among many raters. Psychological Bulletin 1971, 76: 378-382. http://dx.doi.org/10.1037/h0031619

33. Landis JR, Koch GG: The Measurement of Observer Agreement for Categorial Data. Biometrics 1977, 33: 159-174. http://dx.doi.org/10.2307/2529310

34. "Fleiss kappa statistics program," [online] at: https://github.com/dnafinder/Fleiss [Last accessed on October 20 2020]

35. Guyon, I, Weston, J, Barnhill, S, Vapnik V: Gene selection for cancer classificationusing support vector machine. Machine Learning 2002, 46: 389–422. http://dx.doi.org/10.1023/A:1012487302797

36. Kempka, M, Kolowski, W, Warmuth, W, K: Adaptive Scale-Invariant Online Algorithms for Learning Linear Models. 36th Int. Conf. on Machine Learning, Long Beach, CA, PMLR 97, 2019.

37. Shalev-Shwartz, S, Singer, Y, Srebro, N: Pegasos: Primal Estimated Sub-Gradient Solver for SVM. 2007 International Conference on Machine Learning, ICML ’07, 2007, pp. 807–814. http://dx.doi.org/10.1145/1273496.1273598

38. Wei, X: Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent. *CoRR*, abs/1107.2490, 2011.

39. Bifet, A, Gavalda, R: Learning from time-changing data with adaptive windowing. 2007 SIAM international conference on data mining, 2007, pp. 443-448. http://dx.doi.org/10.1137/1.9781611972771.42

40. "The scikit-multiflow implementation of adaptive window kNN," [online] at: https://scikit-multiflow.readthedocs.io/en/stable/api/generated/skmultiflow.lazy.KNNADWINClassifier.html#skmultiflow.lazy.KNNADWINClassifier [Last accessed on October 20 2020].

41. Janowczyk A, Basavanhally A, Madabhushi A: Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology. Comput. Med. Imag. Graph. 2017, 57: 50–61. http://dx.doi.org/10.1016/j.compmedimag.2016.05.003

42. Long L, Shelhamer E, Darell T: Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, Jun 7-12, 2015, pp. 3431-3440. http://dx.doi.org/10.1109/CVPR.2015.7298965

43. Dosovitskiy A, Springerberg JT, Riedmiller M, Brox T: Discriminative unsupervised feature learning with convolutional neural networks. Advances in neural information processing systems (NIPS), Montreal, Canada, December 8-13, 2014 pp. 766–774.

44. Holschneider M, Kronland-Martinet R, Morlet J, Tchamitchian P: A real-time algorithm for signal analysis with the help of the wavelet transform. Proc. Wavelets: Time-Frequency Methods Phase Space, 1989, pp. 289-297.

45. Chen LI, Papandreou G, Kokkinos I, Murphy K, Yuille AL: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convoltion, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 834-848. http://dx.doi.org/10.1109/TPAMI.2017.2699184

46. TensoFlow implementation of the DeepLabv3. [online]: https://github.com/bonlime/keras-deeplab-v3-plus [Last accessed on October 21 2020].

47. TensorFlow White Papers. [online] at: https://www.tensorflow.org/about/bib [Last accessed on October 20 2020].

48. Keras FAQ: Frequently Asked Kears Questions. [online] at: https://keras.io/getting-started/faq/#how-should-i-cite-keras [Last accessed on October 20 2020].

49. Brbić M, Kopriva I: Multi-view Low-rank Sparse Subspace Clustering. Pattern Rec. 2018, 73: 247-268. http://dx.doi.org/10.1016/j.patcog.2017.08.024