Research Article

A U-net based approach to epidermal tissue segmentation in whole slide histopathological images

Kay R. J. Oskal[1] · Martin Risdal[1] · Emilius A. M. Janssen[3] · Erling S. Undersrud[3] · Thor O. Gulsrud[2]

Abstract
Malignant melanoma is a severe and aggressive type of skin cancer, with a rapid decrease in survival rate if not diagnosed and treated at an early stage. Histopathological examination of hematoxylin and eosin stained tissue biopsies under a light microscope is currently the gold standard for diagnosis. However, this manual examination is a difficult and time-consuming task, and diagnosis is often subject to intra- and inter-observer variability. With more pathology departments starting to convert conventional glass slides into digital resources, a Computer Aided Diagnostic (CAD) system that can automate part of the diagnostic process will help address these challenges. It is expected to reduce examination time, increase diagnostic accuracy, and reduce diagnostic variations. An important initial step in developing such a system is an automated epidermis segmentation algorithm, since several important diagnostic factors are within or seen relatively to the epidermis' location. In this paper, we propose a new epidermis segmentation technique built on Convolutional Neural Networks. We trained an U-net based architecture end-to-end, with $\sim$ 380k overlapping high resolution image patches at 512×512 pixels, extracted and augmented from 36 digitized histopathological images from two different clinical sites, to discriminate pixels as either epidermal or non-epidermal. The proposed technique was evaluated on 33 test images, where we achieved a mean Positive Predictive Value at 0.89 ± 0.16, Sensitivity at 0.92 ± 0.1, Dice Similarity Coefficient at 0.89 ± 0.13 and a Matthews Correlation Coefficient at 0.89 ± 0.11, showing a superior performance when compared to existing techniques. Our algorithm also proves to be robust to variations in staining, tissue thickness and laboratory pre-processing.

Keywords CNN · Digital pathology · Melanoma · Neural networks · Semantic segmentation · WSI · U-net

Abbreviations

CAD	Computer aided diagnostic
CNN	Convolutional neural networks
H&E	Hematoxylin and eosin
WSI	Whole slide image
ILSVCRC	ImageNet large-scale visual recognition challenge
FCN	Fully convolutional network
UBC	University of British Columbia
UMch	University of Michigan
ReLU	Rectified linear unit
ELU	Exponential linear unit

1 Introduction

Malignant melanoma is one of the cancer types in Norway with highest increase in incident rate [1], placing Norway amongst the countries in the world with highest melanoma incidence and mortality, when looking at age-standardized rates [2]. In 2017 there were 2222

✉ Kay R. J. Oskal, kay.oskal@norceresearch.no; Martin Risdal, martin.risdal@yahoo.no; Emilius A. M. Janssen, emilius.adrianus.maria.janssen@sus.no; Erling S. Undersrud, erling.sandoy.undersrud@sus.no; Thor O. Gulsrud, thor.o.gulsrud@uis.no | [1]NORCE Norwegian Research Centre AS, Stavanger, Norway. [2]Department of Quality and Health Technology, University of Stavanger, Stavanger, Norway. [3]Department of Pathology, Stavanger University Hospital, Stavanger, Norway.

SN Applied Sciences (2019) 1:672 | https://doi.org/10.1007/s42452-019-0694-y

Received: 8 April 2019 / Accepted: 29 May 2019 / Published online: 7 June 2019

new cases (which is 20 times more that the early 1950s) and 306 dead of the disease, more men than women [3]. Malignant melanoma is among the most aggressive types of skin cancer and if not treated early, the tumor is likely to thicken and progress to a more invasive stage. Here it can invade nearby lymphatic -and/or blood vessels and rapidly spread to other parts of the body. This will cause the 5-year survival rate to drop dramatically from 80–90% (male-female); to only 14–24% depending on cancer stage at time of diagnosis [3]. Detection and accurate diagnosis at an early stage is therefore of the utmost importance.

Histopathological examination of hematoxylin and eosin (H&E) stained tissue biopsies under a light microscope remains the gold standard for diagnosis of malignant melanoma. With this, pathologists have a cellular level view of the disease, and use their deep domain knowledge and experience to asses complex morphological and cytological features of the tissue sample in order to reach a diagnosis. However, the manual evaluation of tissue samples are in many cases complex and therefore a time- and labor-intensive task. In addition, at most pathology laboratories the sheer amount of skin biopsies causes real logistic and personnel challenges [4–6].

Furthermore, since the pathologists' diagnosis is subjective and based on personal experience and bias, this can lead to intra- and inter-observer variability. It has been shown that inter-observer variations of diagnosis sensitivity may range from 55 to 100 percent among 20 pathologist [7].

With more pathology departments being remodeled to *digital pathology*,[1] by converting conventional glass slides into digital resources commonly know as Whole Slide Images (WSIs), a Computer Aided Diagnostic (CAD) system to automate parts of the diagnostic process will help address these challenges. It is expected to reduce examination time, increase diagnostic accuracy, and reduce diagnostic variations.

A digitized H&E stained skin WSI consists of three main parts (i.e. skin layers); epidermis, dermis and subcutaneous tissue, as illustrated in Fig. 1. The epidermis region, highlighted in green, consists of several important diagnostic factors [8]. Therefore, an important initial step in developing a CAD system, is to develop a robust automatic segmentation algorithm to precisely localize this area.

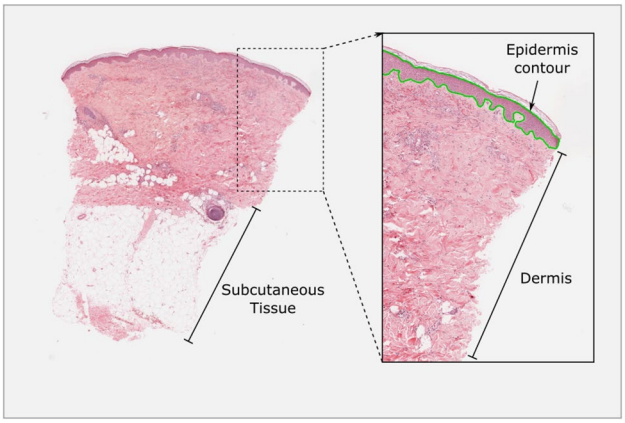


Fig. 1 H&E stained skin biopsy. Illustration of a H&E stained skin biopsy with its three tissue layers: epidermis, dermis and subcutaneous tissue. The green contour of the epidermis area is *manually labeled* and superimposed on the digitized slide

1.1 Related work

Several automated techniques have been proposed for epidermis segmentation. Haggerty et al. [9] proposed a contrast enhancement and thresholding method (henceforth referred to as CET). It performs color normalization on WSIs with 10x magnification. This is followed by global thresholding on contrast enhanced image, which is created from an equally weighted linear combination of the normalized WSIs grayscale-converted image and the blue-yellow component from its CIELAB-converted image.

Lu et al. [10] proposed a technique based on global thresholding and shape analysis (henceforth referred to as GTSA) on WSIs at 40× magnification down-sampled by a factor of 32. This technique first obtains a coarse segmentation by global thresholding on the red channel on the down-sampled image followed by shape analysis. Then, a template matching method is used to enhance the signal of epidermis. A final threshold is then determined by analyzing the probability density function of the response value image.

Xu et al. [11] propose a technique in which an initial coarse segmentation is obtained similar to GTSA followed by a thickness measurement of the obtained epidermis region (henceforth referred to as THM technique). If the coarse segmentation is classified as bad quality (i.e. too thick) a second-pass fine segmentation is performed with a k-means classification algorithm.

Kłeczek et al. [12] propose a technique based on porosity analysis and stain concentration analysis (henceforth referred to as PASC technique) on WSIs from multiple sources down-sampled to 10x magnification. A coarse segmentation is obtained by filling void spaces which are

[1] Image-based information environment which is enabled by computer technology that allows for the management of information gathered from a digital slide.

probable clear cells or desmosomes using a shape criteria and performing global thresholding on density. The next step is to compute hematoxylin (H) and eosin (E) stain concentration and reject blood, stratum corneum and dense collagen by performing global thresholding on E, H and H/E concentration maps. A final refinement is done by rejection highly porous regions, similar to the initial coarse step.

With the exception of the PASC technique, existing techniques makes strong assumptions on staining uniformity and sufficient contrast differences between epidermis and dermis in their approaches. They are all mainly based on global thresholding on a predefined color channel followed by an analysis of shape and area. However, due to inter and intra-variations in staining and tissue thickness, skin appendages and dermal cellular infiltration, these assumptions are often not met. Consequently, these techniques often include large false positive regions within the dermis due to darker components, such as skin appendages and cellular infiltration of nevi cells of lypmhocytes. These color variations also result in general failure for two of the above mentioned methods on almost half of the images when tested by the authors of the PASC technique. This was due to area and/or shape criteria not being met after the global thresholding step. These results are in accordance with our own results, which are presented in Section 5.

In the following pages we propose a robust automatic segmentation technique, built on Convolutional Neural Network (CNN) and the U-net architecture. Our technique does not make any assumption on color channels or contrast. Due to annotated images from multiple sources and the CNN-models ability to learn features that maps each pixel to its respective class, we overcome challenges with anatomical variations and variations that may arise from staining, different slide scanners and software processing.

1.2 Biomedical image segmentation with deep neural networks

Deep neural network is currently the most frequently studied method within the field of machine learning. These learning methods have in the last decade outperformed classical machine learning algorithms with handcrafted features in several fields, including digital image processing. The introduction of CNNs [13] in 1998, a gradient decent based machine learning method with a convolutional architecture, allowed the network to build feature maps directly from annotated training images. However, due to lack of large enough data sets and sufficient computational power, the potential of this technique was not truly shown before the groundbreaking results by Krizhevsky et al. [14]. Their algorithm, AlexNet, was the first deep neural network algorithm to win the ImageNet Large-Scale Visual Recognition Challenge (ILSVCRC) [15], and they won it by a wide margin. Ever since, these types of image analysis competitions are won consistently by deep neural network algorithms.

In the research field of medical imaging, the use of CNNs have gained a growing popularity for various tasks in several modalities, such as radiography, magnetic resonance imaging, ultrasound and computer tomography, among others [16]. Within the relatively new field of digital pathology, analysis of histopathology images also benefits from the multitude of challenges solvable by the use of CNNs. Among these, there has been proposed numerous tasks involving detection, classification and segmentation [16–18].

Regular CNNs map each input image into containing objects from one or several classes (with a possible bounding box describing where the object is), but this results in a too coarse segmentation for the task at hand. Long et al. [19] proposed a solution to this, called Fully Convolutional Network (FCN). This network architecture can be trained end-to-end to obtain a pixel-wise prediction for the entire input image. As of today, most state-of-the-art pixel-wise semantic segmentation networks are based on this approach [20]. FCN based systems can handle input images of arbitrary size, since they are not restricted by any dense layers. However, the giga-pixel nature of histopathological images quickly exceed GPU memory available with current technology. An example WSI with size 40,000 × 60,000 pixels, in RGB color space, would alone need ∼6.7GB of memory. If we use 32 filters in the first layer, the filter activations (i.e. feature maps) would, with single-precision floating points, need roughly 286GB memory. During training, intermediate filter activations are saved at each layer since they are needed for back-propagation. Hence, even for smaller images, the memory requirements will grow fast out of reasonable limits. Two possible solutions are to either down-sample the WSIs resolution with the consequence of losing high resolution details, or a patch-based approach where multiple (possible overlapping) patches are extracted from the full resolution WSI and used as independent input images [21]. One unfortunate effect with the latter approach is the loss of possible global information.

In the development of our system for epidermis segmentation, we have chosen the patch-based approach, which has been common when working with CNNs and full size WSIs. The reason is that, the extensive down-sampling needed to not exceed memory limitations would result in the loss of most of the discriminative cellular level details, which are more important than possible global information.

2 Methods

In this section we present the various data sets used, ground truth labeling and the preparation of patches used for training and testing. Thereafter we present our model's architecture, training, inference and post-processing.

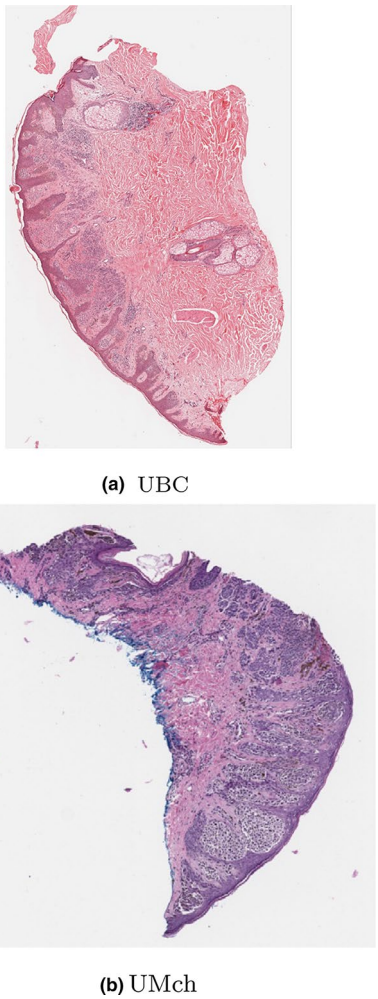
2.1 Data sets

The WSIs used for training and testing our proposed method are all from formalin-fixed paraffin-embedded tissue blocks containing skin biopsies. These are sliced approximately $4\mu m$ thick, stained with H&E using an automated stainer, and are obtained from two different clinical sites.

The first subset of 59 images are publicly available, and obtained from University of British Columbia Virtual Slidebox [22] (henceforth denoted as the UBC data set), scanned with an Aperio ScanScope slide scanner system at an apparent 40x magnification (0.25µm/px) and saved into JPEG format.

The second subset of 10 images are publicly available, and obtained from University of Michigan Virtual Slide Box [23] (henceforth denoted as the UMch data set). These images were scanned with an Aperio ScanScope slide scanner system at an apparent 40× magnification (0.25 µm/px) and saved into JPEG2000 format.

In all, our data set consist of 69 skin WSIs with size ranging between $5000 \times 10,000$ pixels and $48,000 \times 63,000$ pixels. Example WSIs from both of the subsets is shown in Fig. 2.



(a) UBC

(b) UMch

Fig. 2 WSIs from different sources used in this study. This figure shows example WSIs acquired from two different sources

2.2 Ground truth

The boundary of the epidermal area in the WSIs were carefully annotated using Aperio ImageScope, assisted by an expert pathologist specialized in skin diseases at Stavanger University Hospital, and saved as an XML file. Annotations were then converted from XML to a binary image in Matlab. Additionally, foreground masks were obtained by first converting the WSIs to HSV colorspace. Thereafter, we thresholded the H and S channels with Otsu's method, and combined these. Finally, small pixel areas in the background were removed with morphological closing, holes were filled with morphological reconstruction and the boundaries were smoothed with morphological opening (Fig. 3). These masks were used as labels during patch extraction, training, and later as ground truth when evaluating the proposed system.

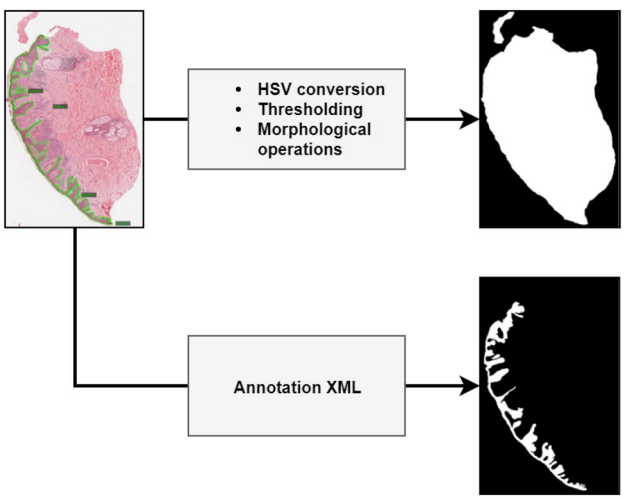


Fig. 3 Foreground and epidermis masks. This figure shortly illustrates how the foreground and epidermis masks are constructed from an annotated WSI

2.3 Training data

For the training of our CNN model, we randomly extracted 384000 overlapping image patches at 512×512 pixels from 36 of the 69 WSIs in our data set. These were chosen from both subsets to represent variations in staining and anatomy. From the total number of image patches, 95% were used to train our model with forward- and backward propagation. The remaining 5% were kept aside as a validation set, used to check if our model were overfitting on the training data, and to fine-tune hyperparameters.

The remaining 33 WSIs were kept aside for testing the model.

2.3.1 Patch extraction

As seen in Fig. 1, the epidermis area comprise of a very tiny part of the total tissue slide. Thus, there is a huge class imbalance between the epidermis region and the rest of the slide. Of the total pixel count in our images, roughly 3.5% of the pixels are within the epidermal area. The remaining pixels are *other tissue* (~29.5%) and *background* (~67%). Due to the insufficient amount of pixels labeled as epidermis, early experiments converged with high accuracy by predominantly classifying pixels as *other tissue* and *background*.

We approached this problem by under-sampling the background and other tissue area, and over-sampling the epidermal area. This was done by first deciding the maximum number of patches from each WSI based on a pixel count in the epidermis ground truth mask. Each new randomly generated patch were extracted by the following criteria:

First, we ensured that the patch was within the boundary of the original WSI. Next, we calculated the euclidean distance between the upper left corner of the new patch and all others, ensuring it being minimum 200 pixels. This was done to spread the patches thoroughly and preventing them from being too overlapping. Then, pixels per class were counted to decide in which category the patch belonged by the following criteria:

- If $> 40\%$ epidermis pixels $\rightarrow$ count as epidermis patch.
- If $> 60\%$ background pixels $\rightarrow$ count as background patch.
- If $> 60\%$ other tissue pixels $\rightarrow$ count as other tissue patch.

If the chosen category were full (i.e. being one third of the maximum number of patches), the patch not being within WSI boarder or it being too close to the other patches, it is discarded and a new patch is randomly generated. This was repeated until all categories were filled. An overview of the described pipeline, for extracting training patches from one WSI, is shown in Fig. 4.

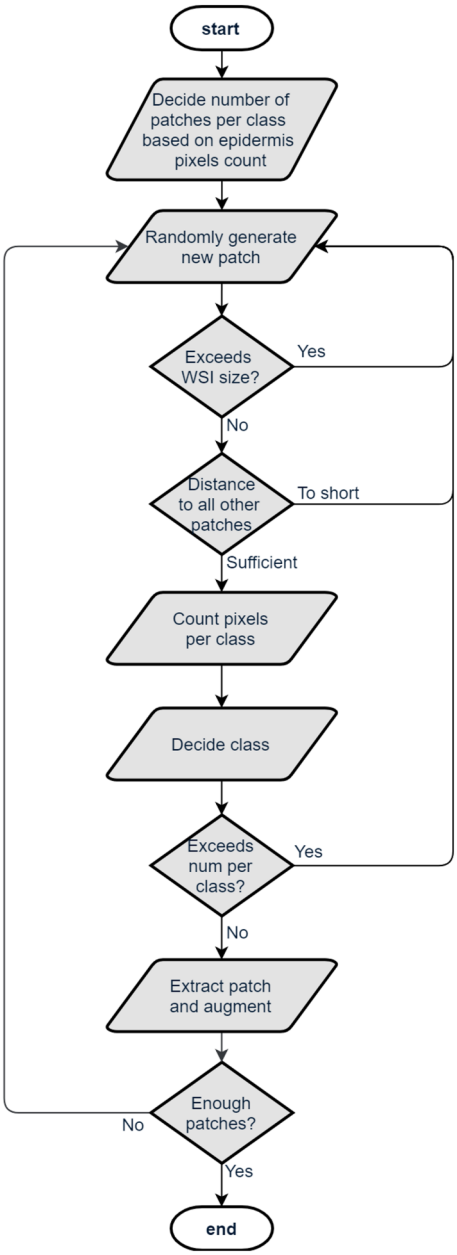


Fig. 4 Patch extraction pipeline. This illustrates how the training patches were extracted from each WSI

This results in a better training set distribution, where we now have 24% of the pixels representing *epidermis* area. Whereas the *other tissues* and *background* areas are represented in 42% and 34% of the total pixel count, respectively. We believe it to be advantageous with an over-representation of *other tissue* pixels because of the high anatomical variance caused by possible hair follicles, sebaceous glands, sweat glands, blood and lymph vessels etc. in the dermis and subcutaneous layers. Additionally, in

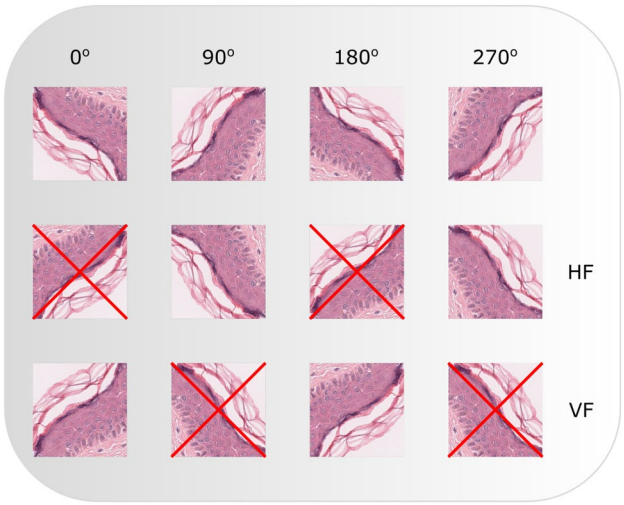


Fig. 5 Patch augmentation. Upper left is the original patch. The remainder are rotated and flipped versions of the original. HF and VF denote horizontal- and vertical flip, respectively. Redundant combinations (crossed out) are excluded

some of the slides, there are different cellular infiltrations within the dermis area, these being lymphocytes, nevus cells and/or tumor cells, which all can vary substantially.

2.3.2 Image augmentation

To further increase the amount of training data, each patch and its corresponding labeled patch were augmented by rotating and flipping the images, as illustrated in Fig. 5. This result in an eight-fold increase of training data.

2.4 Model architecture

In the context of semantic segmentation in medical imaging, the U-net architecture [24], which is an evolution of the FCN architecture, has shown several promising results [25–27]. The first half of the "U" consist of a four level contracting path (encoder), where two 3x3 convolution operations with a rectified linear unit (ReLU) activation are applied, followed by a max-pool down-sampling. This builds a multi-channeled feature map which capture the context in the image, but has a localization trade-off (i.e. the "what" is improved at the expense of the "where"). The localization issue is handled in the second symmetric half, four level expansive (decoder) part of the network. Here, the feature map is up-sampled and concatenated with its corresponding higher resolution feature map from the contracting path, which combines context information with precise localization. Additional convolutional operations are, at each level applied to the concatenated feature maps, which assembles a more precise output. A final 1x1 convolution is then used to map the last feature map to its desired class.

Our adaption of this architecture, shown in Fig. 6, has some modifications. We halved the number of feature channels to 32 - 64 - 128 - 256 - 512, thus drastically reducing the memory required for filter weights and to store all intermediate filter activations which is needed for back-propagation. With the original number of feature channels and a batch size of eight, the model would require ~ 25GB GPU memory during training, exceeding the limits on every GPU on the market. However, with the number of feature channels halved, model requirement is ~ 12.7GB, which is doable on higher end GPUs. At every convolutional layer, except the last 1x1, we use zero-padded convolutions. Hence, no cropping operations are needed before the concatenation of feature maps. In addition, this preserves the spatial dimensions of the input to the output. Furthermore, the ReLU activations has been switched with the Exponential Linear Unit (ELU) [28] activations:

$$f_{ELU}(x) = \begin{cases} \alpha(e^x - 1) & for\ x < 0 \\ x & for\ x \geq 0 \end{cases} \quad (1)$$

where α is a hyperparameter that controls the value to which ELU saturates for negative inputs. According to the original ELU article, it will provide much faster learning and a better generalization performance compared to ReLU. To further reduce the training time, we use *batch normalization* [29] after each convolutional layer, which also provides some regularization effect to the model. Additional regularization is done by incorporating *dropout* [30] at each level in both contracting and expansive path, unlike the original architecture which only has it in the last two levels of the contractive path. Due to the high resolution images, the original filter size of 3×3 has been increased to 5×5, which increases the models receptive field. Thus more contextual information and finer details are expected to be captured.

2.5 Training and inference

To focus on the classification of the epidermal area only, the classes *other tissue* and *background* were combined as *non-epidermal area*, hence making this a binary classification problem. The model proposed in this study maps $512 \times 512 \times 3$ RGB images to $512 \times 512 \times 1$ images with floating-point pixel values from 0 to 1, where each pixel value represent the predicted probability of that given pixel belonging to the epidermal area. Furthermore, to produce the final output mask, these predicted images were post-processed to result in a proper binary mask.

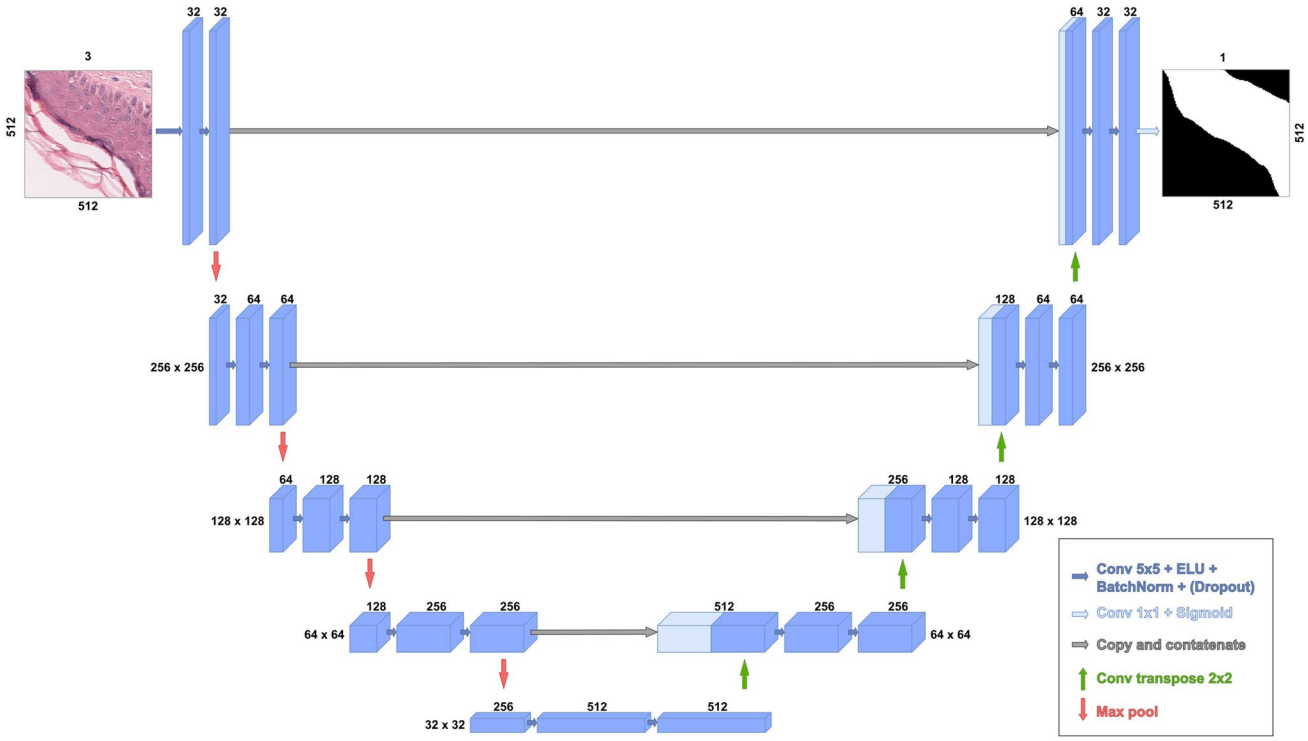


Fig. 6 Model architecture. Model architecture adapted from Olaf Ronnebergers U-net: Convolutional Networks for Biomedical Image Segmentation [24]. Each blue box correspond to a multi-channel feature map. The number of channels are shown over the boxes. Light blue boxes are copied feature maps from the encoder part of the network. Finally, the arrows denote the different operations

2.5.1 Training

We trained the model over 15 epochs with a batch-size of eight on training data containing 364800 image patches. Prior to the training, all kernels were initialized using *He's uniform* initialization [31]. When training a neural network, a stochastic gradient-based optimization is used to update the parameters in the model. Here we used an *Adam* optimizer [32], which computes an adaptive learning rate to further speed up the training. The learning rate was initially set to $\alpha_{init} = 2 \times 10^{-3}$, and was reduced by a factor of 2 if loss didn't improve during the last two epochs. To prevent overfitting, dropout with a drop rate of 0.2 were used on the first convolutional layer at each level, both in the contractive and the expansive path.

2.5.2 Loss function

At the end of the forward-pass for each mini-batch, training loss were derived from the sigmoid activation output at the last layer

$$\hat{y} = \sigma(x) = \frac{1}{1 - e^{-x}} \tag{2}$$

and by calculating its binary cross entropy loss

$$\mathcal{L}(\hat{y}, y) = -(ylog(\hat{y}) + (i - y)log(i - \hat{y})) \tag{3}$$

This loss was then back-propagated throughout the network, updating each parameter.

2.5.3 Inference

To infer the segmentation, non-overlapping patches were extracted from a full size WSI and fed forward into the model, which generates a epidermis probability map. These were then reassembled into an image equal in size as its original WSI. To reduce checkerboard artifacts caused by the edges of each patch, we increased the patch-size in the inference to 4096×4096. Since FCN based systems doesn't include any dense layer, the input size of the network isn't restricted by other than available GPU memory, and that the dimension of the input image must be dividable by 2^n where $n = \{1, 2..., N\}$ and N is the depth of the network.

2.5.4 Post-processing

To reduce computational complexity in the post processing stage, the output probability map from the inference

Table 1 Segmentation performance on the training set (n = 36)

Method	#failed	Mean value				Median value				SD			
		$\mathcal{A}_{PPV}$	$\mathcal{A}_{SEN}$	$\mathcal{A}_{DSC}$	$\mathcal{A}_{MCC}$	$\mathcal{A}_{PPV}$	$\mathcal{A}_{SEN}$	$\mathcal{A}_{DSC}$	$\mathcal{A}_{MCC}$	$\mathcal{A}_{PPV}$	$\mathcal{A}_{SEN}$	$\mathcal{A}_{DSC}$	$\mathcal{A}_{MCC}$
CET	0	0.31	**0.97**	0.41	0.46	0.22	**0.99**	0.36	0.43	0.26	0.06	0.27	0.24
GTSA	20	0.66	0.55	0.53	0.55	0.76	0.64	0.54	0.59	0.26	0.36	0.28	0.27
THM	13	0.69	0.39	0.42	0.45	0.73	0.20	0.31	0.34	0.20	0.35	0.30	0.27
PASC	0	0.50	0.89	0.59	N/A	0.52	0.97	0.61	N/A	0.27	0.21	0.26	N/A
Proposed	0	**0.84**	**0.97**	**0.86**	**0.87**	**0.96**	0.98	**0.97**	**0.96**	0.28	0.02	0.24	0.21

were decimated by a factor of four. This also made it so the output size matched the existing techniques, which eases the performance comparison. The probability map were binarized by first smoothing the image with an 11×11 averaging kernel, before the image was thresholded using using Otsu's method [33]. Small objects with areas smaller than 20,000 pixels were removed to eliminate false positives. To finalize the epidermis mask, morphological opening with a disk-shaped kernel with radius of eight, were applied to smooth the boundary. The kernel sizes used for averaging and morphological opening and the size threshold for pixel areas to be removed, were obtained empirically.

2.6 Implementation details

All experiments were conducted using Keras [34] deep learning library with Tensorflow backend [35] in Python 3.5. Both training and inference were done on a NVIDIA Tesla P100 GPU computing processor with 12GB of memory.

3 Results

In this section we present comparative segmentation results by our proposed method and existing techniques.

3.1 Evaluation metrics

The segmentation results are compared to the ground truth by four area based metrics:

- *Positive Predictive Value* ($\mathcal{A}_{PPV}$), measures how precise the segmentation mask is within the boundary of the ground truth.
- *Sensitivity* ($\mathcal{A}_{SEN}$), measures how large part of the ground truth mask is covered by the segmentation mask.
- *Dice Similarity Coefficient* ($\mathcal{A}_{DSC}$), is the harmonic average of the $\mathcal{A}_{PPV}$ and $\mathcal{A}_{SEN}$.
- *Matthews Correlation Coefficient* ($\mathcal{A}_{MCC}$), is a balanced measure that uses all the four classes of the confusion matrix in its computation.

As a single score, $\mathcal{A}_{PPV}$ and $\mathcal{A}_{SEN}$ won't provide an accurate measure of the actual performance, and must therefore be evaluated as a pair. $\mathcal{A}_{DSC}$ score is widely used, but can be misleading when working with large class imbalance. This is the case when classifying a relatively small epidermal area in a large image. Hence, the score strongly depend on which class is defined as the positive class. When calculating the $\mathcal{A}_{MCC}$ score, both the positive and negative elements are considered and is therefore independent of which class is defined as what. Thus, it's claimed to be the most informative single score performance metric in binary classification [36]. The four evaluation metrics are calculated as follows:

$$\mathcal{A}_{PPV} = \frac{TP}{TP + FP} \tag{4}$$

$$\mathcal{A}_{SEN} = \frac{TP}{TP + FN} \tag{5}$$

$$\mathcal{A}_{DSC} = 2 \cdot \frac{\mathcal{A}_{PPV} \cdot \mathcal{A}_{SEN}}{\mathcal{A}_{PPV} + \mathcal{A}_{SEN}} \tag{6}$$

$$\mathcal{A}_{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

3.2 Quantitative results

To evaluate the efficiency of our proposed system, its performance is compared to the segmentation techniques CET, GTSA, THM and PASC presented in the Related Work section. The CET, GTSA and THM techniques all have several key parameters, which are set in accordance with their original work. The PASC results are obtained from the original authors online research page [37].

Metrics defined in Eqs. 4–7 were calculated from the resulting segmentation maps produced by each method and are summarized in Tables 1 and 2. Highlighted in bold

Table 2 Segmentation performance on the test set (n = 33)

Method	#failed	Mean value				Median value				SD			
		$\mathcal{A}_{PPV}$	$\mathcal{A}_{SEN}$	$\mathcal{A}_{DSC}$	$\mathcal{A}_{MCC}$	$\mathcal{A}_{PPV}$	$\mathcal{A}_{SEN}$	$\mathcal{A}_{DSC}$	$\mathcal{A}_{MCC}$	$\mathcal{A}_{PPV}$	$\mathcal{A}_{SEN}$	$\mathcal{A}_{DSC}$	$\mathcal{A}_{MCC}$
CET	0	0.35	**0.99**	0.47	0.52	0.34	**0.99**	0.50	0.55	0.22	0.01	0.25	0.22
GTSA	21	0.73	0.31	0.39	0.42	0.81	0.19	0.29	0.35	0.27	0.31	0.33	0.31
THM	17	0.69	0.38	0.45	0.47	0.72	0.26	0.39	0.42	0.20	0.32	0.28	0.27
PASC	0	0.65	0.84	0.68	N/A	0.72	0.96	0.77	N/A	0.25	0.26	0.23	N/A
Proposed	0	**0.89**	0.92	**0.89**	**0.89**	**0.95**	0.96	**0.94**	**0.93**	0.16	0.10	0.13	0.11

are the technique which achieved the highest score on the given performance metric. The *#failed* column present the number of WSIs for which a given method failed (i.e. not a single pixel is segmented).

The mean, median and standard deviation in the tables are all calculated from the *non-failed* images, to emphasize on the respective techniques performance on images where a segmentation result were obtained.

As shown in bold in the tables, our proposed system provides an overall superior performance according to almost all metrics. The only metric where other methods supersede our system is the sensitivity achieved by CET. However as explained, sensitivity alone isn't enough to evaluate the techniques overall performance, since sensitivity doesn't account for false positive areas (e.g. an image where all pixels are segmented as epidermis, will get a sensitivity score of 1). If we look at CET's other metrics, we observe that they are significantly lower than those of our method. As mentioned, this is mainly due to incorrect classification of abundant non-epidermis pixels as epidermis.

As a single score metric, $\mathcal{A}_{DSC}$ and $\mathcal{A}_{MCC}$ are much more descriptive. To emphasize the differences according to these metrics, the proposed method achieves a $\mathcal{A}_{DSC}$ score higher than 0.9 in 25 of 31 test WSIs, whereas none of the other techniques scored this high on more than one of the images in the test set. This superior performance is also shown when inspecting $\mathcal{A}_{MCC}$, where our method scored higher than 0.9 in 23 images. These values are also supported if we inspect the mean and median from $\mathcal{A}_{DSC}$ and $\mathcal{A}_{MCC}$. Additionally, the robustness of our method never lead to failure on any of the images, whereas both GTSM and THM fails on around half of them.

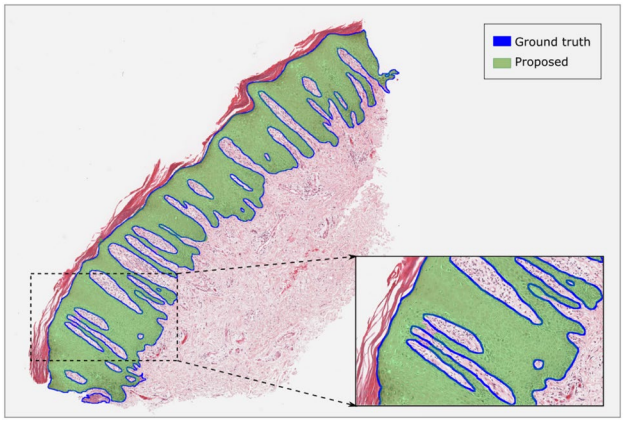


Fig. 7 Example representing majority of test slides. A chosen example representing the 25 slides which received a $\mathcal{A}_{DSC} > 0.9$. This slide in particular scored *0.97* on all performance metrics

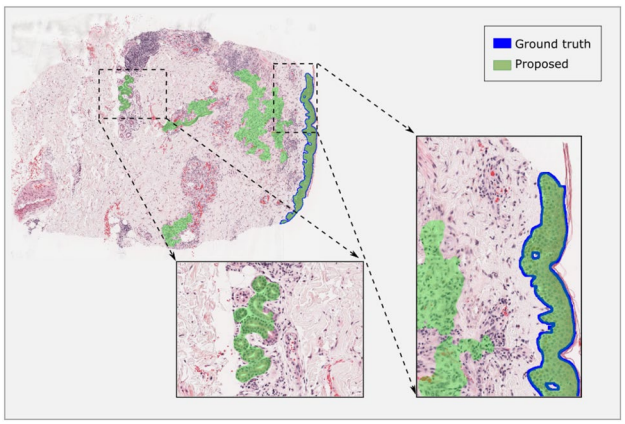


Fig. 8 Example slide with false positives. An example representing the six slides which received a $\mathcal{A}_{DSC} < 0.9$ due to false positives. This slide in particular achieved $\mathcal{A}_{PPV} = 0.25$, $\mathcal{A}_{SEN} = 0.98$, $\mathcal{A}_{DSC} = 0.40$, $\mathcal{A}_{MCC} = 0.49$

3.3 Qualitative evaluation

Qualitative segmentation results from our proposed method are illustrated in Figs. 7, 8, and 9. Overall, the accuracy of our proposed technique is seen on a majority of the images in our data set. Here, most of the true epidermal area is correctly predicted, with almost negligible amount of non-epidermal pixels predicted as epidermis. The slide shown in Fig. 7 is an example representing the 25 WSIs that show this accuracy. In this case we can observe that our proposed technique segment an area almost identical to the ground truth mask. Similar performance is also seen in the remainder of these mentioned slides. However, for the reasons seen below, the proposed technique didn't perform satisfactory on some of the images.

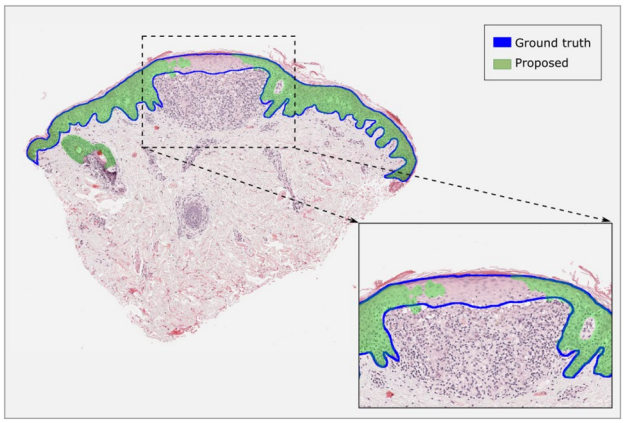


Fig. 9 Example slide where the model were unable to segment whole epidermis area. A chosen example representing the four slides which received a dice score below 0.9 due to false negatives in the epidermal area. This slide in particular also have a false positive area which contribute to lowering the score

Of the six images receiving a $\mathcal{A}_{DSC}$ score lower than 0.9, two had low positive prediction value due to a noticeable amount of false positives, as the example shown in Fig. 8. The false positive areas seen in these images are due to areas contain hair follicles, sweat glands and sebacceous glands which all contain epithelial cells. These are probably mistaken as the similar keratinocytes, which also is epithelial cells, and the major component of the epidermal area. Some areas containing lymphocytes were also wrongly classified in these images.

The remaining four with low $\mathcal{A}_{DSC}$ were due to a low sensitivity score, and inspection showed that the model was not able to segment the whole epidermal area in these cases. One of these WSIs is shown in Fig. 9. The reasons as for why our model was unable to fully segment the epidermal area in these images, varies from image to image. One image has a piece of folded tissue in the epidermis area, which our system didn't recognize. Another image comprise of an epidermis broken up into several parts, where some of these were deemed as too small and removed in the post-processing stage. The last two images suffered from areas abundant with clear cells within the epidermis, resulting in lower values in the probability map outputted from our deep neural network model. Consequently, these were removed in the thresholding step of the post-processing. Hence, the issue in these four cases partially arise in the post-processing stage, which in these images removed too much of the actual epidermal area.

Performance metrics and segmentation results for each individual WSI can be found online.[2]

[2] http://bit.ly/u-net-epidermis

Table 3 Examples of inference times compared to WSI size

WSI size (pixels)	Inference time (s)
6,125×7754	22.5
17,111×17,145	136.5
26,536×24,406	238
39,252×32,831	505.3
63,002×48,524	1075.1

3.4 Computational complexity

Due to the varying size of the WSIs, the inference time will vary proportionally to the images size. On average, the inference time for a full-size WSI in our data set, was 138 s. For each individual 4096×4096 patch, the inference time is $\approx 4.3s$. Inference times on a selection of WSIs are shown in Table 3.

4 Discussion

The aim of our study was to develop, train and test an automated method for segmenting epidermal area in digitized H&E stained skin WSI with convolutional neural networks.

Both the GTSA and THM techniques make strong assumptions about color and contrast in their approaches, and are both mainly based on global thresholding on the red channel of an RGB image in addition to area and shape analysis. Due to nuances in anatomy, inter- and intra-variations in staining and tissue thickness, these assumptions are in many cases not met. Regions containing stratum corneum, infiltration of lymphocytes or nevi cells, skin appendages and inking of biopsy edges are often regions highlighted using Otsu thresholding. Thus the two above mentioned methods, using this as a coarse segmentation step, often fail on their area and/or shape criteria.

All compared techniques, including the cases for which GTSA and THM didn't fail, score relatively low on Positive Predictive Value. This is mainly due to regions within dermis being infiltrated with lymphocytes and/or nevi cells, or due to the specimen containing numerous skin appendages, which will appear as low-intensity areas similar to the epidermis. Consequently, making the global thresholding incapable of discriminating these areas as non-epidermal. The THM method tries to limit false positive areas by measuring the epidermis thickness followed by an additional fine segmentation using a k-means algorithm. The PASC method reduces the number false positives by rejecting regions within dermis, based on their position relative to the lesion's boundary. However, it fails to reject dense

connective tissue and cellular infiltration close to the lesion's boarder.

The proposed technique does not make any assumptions and only relies on the models ability to *learn* features from training data from several different sources. This allows our model to map each WSI pixel to its respective class, without ever failing on any of the images. However, as with all existing methods, our method sometimes suffer from false positive predicted pixels. But, it doesn't happen as often and at a much smaller scale. Overall, the proposed technique is more robust and provides a more accurate segmentation when compared to existing techniques.

A drawback with the proposed technique, is that high computationally complexity lead to long inference time. However, the inference time is believed to decrease considerably with some optimization in regard to coding. Additionally, the patches which only consist of background could be excluded from the neural network pipeline. Thus, reducing the inference time further.

5 Conclusion

This paper presents a new method for segmenting the epidermal areas in hematoxylin and eosin stained whole slide histopathological images, using Convolutional Neural Network and a U-net based architecture. The proposed method is trained end-to-end on 384 000 image patches extracted from 36 different whole slide images from two different sources, with binary cross entropy loss. Furthermore, the predicted segmentation maps generated by our CNN model, are post-processed to produce proper binary segmentation masks. The proposed technique was evaluated both on the training WSIs and on the 33 test image, where it achieved satisfactory results in both instances. On the training data it achieved an overall mean Positive Predictive Value at 0.84 ± 0.28, Sensitivity at 0.97 ± 0.07, Dice Similarity Coefficient at 0.86 ± 0.24 and a Matthews Correlation Coefficient at 0.87 ± 0.21. On the 33 test images, our method achieved a mean Positive Predictive Value at 0.89 ± 0.16, Sensitivity at 0.92 ± 0.1, Dice Similarity Coefficient at 0.89 ± 0.13 and a Matthews Correlation Coefficient at 0.89 ± 0.11. In a vast majority of the images, our method was able to correctly predict the full epidermal area with very few false positive areas. However, in some WSIs there were too many false positive pixels, thus lowering the mean results. This was mainly due to cellular infiltration and/or skin appendages within the dermis area, incorrectly classified as epidermis. There was also a few cases where the system showed inability to segment the complete epidermal area.

Our method proves to be robust to variations in staining, tissue thickness and laboratory pre-processing. With a larger training set and with more diverse images, the proposed method is believed to overcome the issues seen in the images struggling with cellular infiltration and skin appendages. Additional color augmentation of the training patches is also believed to help rectify issues seen in these cases. With these issues solved, we strongly believe our method could be good a foundation when developing a CAD system assisting pathologists during their diagnostic process.

Author's contribution KRJO implemented and tested the algorithms presented in the manuscript, and prepared the first draft of current manuscript. MR and TOG participated in the design and coordination of this study, and revised the the manuscript. EAMJ proposed the main project for which this manuscript is a part of and is KRJO's main contact at Stavanger University Hospital. ESU provided expert advice and contributed with the annotation of the epidermal area on all WSIs used in this study. All contributors have given final approval of the version to be published.

Compliance with ethical standards

Availability of data and materials The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate The images used in this study are available publicly online, therefore no ethics approval is needed.

Conflict of interest The authors declares that they have no conflict of interests.

References

1. Cancer in Norway 2016; 2017. https://www.kreftregisteret.no/Generelt/Publikasjoner/Cancer-in-Norway/cancer-in-norway-2016/
2. Farlay J, Ervik M, Lam F, Colombert M, Mery L, Piñeros M, et al (2018) Global cancer observatory: cancer today; Accessed Sept 2018. Available from: http://gco.iarc.fr/today/home
3. National cancer registry (2018) https://www.kreftregisteret.no/Generelt/Fakta-om-kreft/Foflekkreft/
4. Robboy SJ, Weintraub S, Horvath AE, Jensen BW, Alexander CB, Fody EP et al (2013) Pathologist Workforce in the United States: I. Development of a predictive model to examine factors influencing supply. Arch Pathol Lab Med 137(12):1723–1732. https://doi.org/10.5858/arpa.2013-0200-OA
5. Report of the National Task Force on Medical Staffing;. Accessed Sept 28, 2018. Available from: http://health.gov.ie/wp-content/

uploads/2014/03/Report-of-the-National-Task-Force-on-Medical-Staffing-Hanly-report.pdf
6. McBride M (2018) Severe shortage of pathologists threatens Israel's health system—especially cancer testing;. Accessed Sept 28. Available from: https://www.darkdaily.com/severe-shortage-of-pathologists-threatens-israels-health-system-especially-cancer-testing-51711/
7. Brochez L, Verhaeghe E, Grosshans E, Haneke E, Piérard G, Ruiter D et al (2002) Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions. J Pathol 196(4):459–466. https://doi.org/10.1002/path.1061
8. Urso C, Rongioletti F, Innocenzi D, Saieva C, Batolo D, Chimenti S et al (2005) Interobserver reproducibility of histological features in cutaneous malignant melanoma. J Clin Pathol 58(11):1194–1198
9. Haggerty JM, Wang XN, Dickinson A, O'Malley CJ, Martin EB (2014) Segmentation of epidermal tissue with histopathological damage in images of haematoxylin and eosin stained human skin. BMC Med Imaging 14(1):7. https://doi.org/10.1186/1471-2342-14-7
10. Lu C, Mandal M (2015) Automated analysis and diagnosis of skin melanoma on whole slide histopathological images. Pattern Recognit 48(8):2738–2750
11. Xu H, Mandal M (2015) Epidermis segmentation in skin histopathological images based on thickness measurement and k-means algorithm. EURASIP J Image Video Process 1:18. https://doi.org/10.1186/s13640-015-0076-3
12. Kłeczek P, Dyduch G, Jaworek-Korjakowska J, Tadeusiewicz R (2017) Automated epidermis segmentation in histopathological images of human skin stained with hematoxylin and eosin. In: Proceedings of SPIE. vol 10140
13. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
14. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems 25. Curran Associates, Inc.; pp 1097–1105. Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
15. ImageNet;. http://image-net.org/
16. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
17. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al (2016) Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 6(1). Available from: https://www.ncbi.nlm.nih.gov/pubmed/27212078
18. Angel CR, Ajay B, Fabio G, Hannah G, Michael F, Shridar G, et al (2014) Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks, vol 9041; p 15. Available from: https://doi.org/10.1117/12.2043872
19. Shelhamer E, Long J, Darrell T (2016) Fully convolutional networks for semantic segmentation. CoRR. arXiv:abs/1605.06211
20. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, et al (2018) Understanding Convolution for Semantic Segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV); pp 1451–1460
21. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH (2016) Patch-based convolutional neural network for whole slide tissue image classification. In: The IEEE conference on computer vision and pattern recognition (CVPR)
22. University of British Columbia. UBC Virtual Slidebox; 2013. Accessed Nov 2017. Available from: http://histo.anat.ubc.ca/PATHOLOGY/Anatomical%20Pathology/DermPath/
23. University of Michigan. University of Michigan Virtual Slide box; 2013. Accessed Nov 2017. Available from: https://www.pathology.med.umich.edu/slides/search.php?collection=Andea&dxview=show
24. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical image computing and computer-assisted intervention - MICCAI 2015. Springer, Cham, pp 234–241
25. Novikov AA, Major D, Lenis D, Hladuvka J, Wimmer M, Bühler K (2017) Fully convolutional architectures for multi-class segmentation in chest radiographs. CoRR. arXiv:abs/1701.08816
26. Dong H, Yang G, Liu F, Mo Y, Guo Y (2017) Automatic brain tumor detection and segmentation using U-net based fully convolutional networks. Commun Comput Inf Sci Med Image Underst Anal 723:506–517
27. Norman B, Pedoia V, Majumdar S (2018) Use of 2D U-net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR Imaging data to determine relaxometry and morphometry. Radiology 288(1):177–185
28. Clevert D, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by exponential linear units (ELUs). CoRR. arXiv:abs/1511.07289
29. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR. arXiv:abs/1502.03167
30. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
31. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. CoRR. arXiv:abs/1502.01852
32. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. CoRR; arXiv:abs/1412.6980
33. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66
34. Chollet F, et al. (2015) Keras: the python deep learning library; Available from https://keras.io
35. Martín A, Ashish A, Paul B, Eugene B, Zhifeng C, Craig C, et al (2015) TensorFlow: large-scale machine learning on heterogeneous systems; software available from tensorflow.org. Available from https://www.tensorflow.org/
36. Chicco D (2017) Ten quick tips for machine learning in computational biology. BioData Min 10(1):35. https://doi.org/10.1186/s13040-017-0155-3
37. Kłeczek P (2017) Automated epidermis segmentation in histopathological images of human skin stained with hematoxylin and eosin; Accessed Sept 2018. Available from http://home.agh.edu.pl/~pkleczek/dokuwiki/doku.php?id=research:skin:epidermis_segmentation_spie2017:epidermis_segmentation_spie2017