2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

SEGMENTATION AND CLASSIFICATION OF MELANOMA AND NEVUS IN WHOLE SLIDE IMAGES

Mike van Zon[1] & Nikolas Stathonikos[2], Willeke A.M. Blokx[2], Selim Komina[3], Sybren L.N Maas[2], Josien, P.W. Pluim[1], Paul J. van Diest[2], Mitko Veta[1],

[1]Eindhoven University of Technology, The Netherlands
[2]University Medical Center Utrecht, Department of Pathology, The Netherlands
[3]University "Ss Ciril and Methodius", Institute of Pathology, Skopje, R. of North Macedonia

ABSTRACT

The incidence of skin cancer cases and specifically melanoma has tripled since the 1990s in The Netherlands. The early detection of melanoma can lead to an almost 100% 5-year survival prognosis dropping drastically when detected later. Studies show that pathologists can have a discordance reporting of melanoma to nevi up to 14.3%.
An automated method could help support pathologists in diagnosing melanoma and prioritize cases based on a risk assessment.
Our method used 563 whole slide images to train and test a system comprising of two models that segment and classify skin sections to melanoma, nevus or negative for both. We used 232 slides for training and validation and the remaining 331 for testing. The first model uses a U-Net architecture to perform a semantic segmentation and the output of that model was used to feed a convolution neural network to classify the WSI with a global label. Our method achieved a Dice score of 0.835 ± 0.08 on the segmentation of the validation set and a weighted F1-score of 0.954 on the independent test dataset. Out of the 176 melanoma slides, the algorithm managed to classify 173 correctly. Out of the 62 nevi slides the algorithm managed to correctly classify 57.

Index Terms— Digital pathology, deep learning, histopathology, melanoma,

1. INTRODUCTION

Melanoma is an aggressive form of skin cancer, with growing incidence rates worldwide. In The Netherlands the incidence has more than tripled since the 1990s, with 6588 diagnosed cases in 2016[1]. Although the 5-year survival of melanoma has been steadily increasing over the last decades, with a current survival rate of 91.5%, early detection is still crucial. Most cases are detected in early stage I when surgical removal of the lesion is often sufficient, leading to a corresponding 5-year survival prognosis of almost 100%[1]. However, if melanoma is detected in later stages the 5-year survival drastically drops, to 75%, 62% and 19% in stage II, III and IV respectively[1].

In early stages melanoma can look very similar to a common moles (nevus) on the dermatological level, but can be distinguished by clinical features such as larger size, irregular shape and border and varying colors. Upon suspicion the lesion will be, when possible, completely excised and sent for histological analysis. This forms the gold standard for the pathological diagnosis of melanoma and is critical in guiding the therapeutic approach. However, on the histopathological level differentiating between benign nevus and melanoma can be challenging too, and lack of agreement among dermatopathologists on the diagnosis has been observed in multiple studies [2]–[6] with a study reporting a discordance of 14.3% between the diagnosis of melanoma and nevi[5]. There have been several studies which have analyzed skin sections mainly focused on either tumor segmentation or melanoma classification. Andres et al.[7] used a random forest classifier to detect melanoma tumor regions and subsequently identified and classified mitotic figures, aimed at supporting pathologists in diagnosing melanoma. Xu et al.[8] proposed a method using support vector machines (SVM) to directly classify skin WSIs as melanoma, nevus or negative. Using cross-validation they obtained an accuracy of 98% on 66 WSI.

In this work, we present a deep learning framework that combines lesion segmentation and slide-based classification. Our model creates a risk map overlay that highlights areas of predicted melanoma, as well as areas of predicted nevus. Secondly, our model provides a classification of whether a whole slide image (WSI) contains melanoma, nevus, or is negative for both.

2. DATASET

The dataset consisted of a total of 563 skin excision and biopsy slides retrieved from the diagnostic pathology archive of the University Medical Center Utrecht, The Netherlands. We selected cases diagnosed as melanoma, common nevus or negative for both (normal skin), and subsequently selected one slide per case stained with hematoxylin and eosin (H&E). Out of the 563 slides, 232 were manually annotated by an experienced pathologist (SK) in dermatopathology as a subspecialty and a senior pathology resident (NM) and then checked and confirmed by an experienced dermatopathologist (WB) that specializes in melanoma. The pathologists used pixel level annotations segmenting the lesions and annotating structures like inflammation and

regression. If there were multiple sections on a glass slide, only one representative section was annotated. Out of the 202 slides in the training set there were 83 slides classified as melanoma, 97 as nevus and 52 as negative. Twenty percent of the slides was reserved as a validation set. The training set was scanned using two scanners, a Philips Ultra Fast Scanner at ×40 magnification with 0.25 μm pixel size and a Hamamatsu Nanozoomer 2.0 XR using ×40 magnification with 0.23 μm pixel size. Thirty extra nevus slides were then scanned on the Hamamatsu scanner after we noticed that the staining on the glass slides had faded. This was done to ensure that the Hamamatsu dataset had comparable number of WSI per class. The resulting dataset for the Hamamatsu scanner was 97 nevus WSI, 83 melanoma WSI and 52 negative WSI. The rest of the 563 slides – 331 WSI - were reserved for testing using only the slide level classification. The slide level classification was obtained from the diagnosis in the pathology reports. The dataset was scanned using both scanners to ensure that the model is not scanner dependent and would learn features that were not tied to a specific scanner color output.

Table 1: Number of unique WSIs divided over training, validation and test set. The * + symbols denote the WSI that were scanned on both scanners.

Diagnosis	Scanner	Training	Validation	Test
Melanoma	Philips	66*	17*	176
	Hamamatsu	66*	17*	0
Nevus	Philips	53	14	62
	Hamamatsu	24	6	0
Negative	Philips	41+	11+	93
	Hamamatsu	41+	11+	0

3. METHOD

The goal was to classify a histopathology WSI of a skin biopsy as either melanoma, nevus or negative for both. In order to achieve this, we proposed a two-model framework. The first model performs a pixel-wise classification between the three classes (melanoma, nevus and negative) to end up with a semantic segmentation over the whole slide. The second model then processes the WSI together with the semantic segmentation to obtain the slide-level classification.

We trained the U-Net using 512×512 patches sampled from ×10 magnification (1μm/pixel) of the annotated areas in our training set. We first started by sampling 1,000×1,000 areas and then sampling 512×512 patches from them. We applied data augmentation on the patches during training with random rotation, color channel shifts, contrast augmentation, Gaussian blurring to compensate for slide blurring artifacts as well as random zooming to compensate for the difference in pixel size between the scanners.

All model were implemented in Keras using the Tensorflow backend. The U-Net[9] model uses categorical cross-entropy for loss, and is trained using the ADAM optimizer with decoupled weight decay. A batch size of 6 is used and training is continued until the validation loss does not increase over 12,500 consecutive iterations.

The initial version of the U-Net was quite prone to false positives on more difficult, underrepresented, areas like glands, damaged tissue and ink. To improve model performance on these regions, we performed a hard-negative mining step[10] to obtain a better balanced training set.

The hard-negative set is obtained by running the initial model over the bounding boxes of the manual annotations. The resulting predictions were split up in patches of 1,000×1,000 pixels (with 50% overlap), and divided in bins based on their accuracy compared to the ground truth. The hard-negative dataset is obtained by taking 40 patches evenly distributed across bins for each slide and saving them to memory. The hard-negative dataset replaced the previous method where 1,000×1,000 areas were randomly sampled from annotated regions while the data augmentation process remained the same. Training of the U-Net model was continued, this time using stochastic gradient descent instead of ADAM as the optimizer, with the hard-negative data until the validation loss did not increase for 5,000 iterations.

The final trained U-Net was applied in a patch-based fashion over the slide to produce a whole-slide semantic segmentation. We first apply a simple tissue detection algorithm by converting the image from RGB to HSV and then calculating the Otsu's threshold on the H and S channel and finally combining them in to a binary mask. Only positive areas were evaluated by the U-Net reducing processing time substantially. Since the U-Net was fully convolutional it is possible to evaluate on arbitrary sized patches, so to further speed up computation 1024×1024 patches are used during evaluation. To prevent border artifacts between patches, we evaluated with a stride of 854 pixels.

The second step was to classify the entire slide as either melanoma, nevus or negative for both utilizing the semantic segmentation obtained from the U-Net. The second model is necessary because the semantic segmentation will very often misclassify small structures such as sweat glands or blood vessels as melanoma or nevi so it would produce a very high number of false positives if we were to accept the mere presence of these structures as evidence of melanoma or nevi. The CNN model receives larger areas of the image as input and based on the context, it can determine if the segmentation corresponds to a global label for the slide.

We applied a patch-based approach using a three-layer CNN followed by a fully connected layer with a softmax as an output layer. The input of the model is a 256×256 6-channel patch (8 μm/pixel, ×1.25 magnification) consisting of the tissue patch together with its corresponding semantic segmentation. The output of the CNN is a softmax probability of this area over the 3 classes (melanoma, nevus and negative).

Training data for the CNN was obtained by extracting the 256×256 patch around a random location within the manual annotation, similar to the U-Net training process. The training and validation split was identical to that used for training the U-Net. Training was done using categorical cross-entropy loss and the ADAM optimizer with a batch size was set to 12 and training was halted if the validation loss did not increase over 250 consecutive iterations.

To classify the slide, we split up the WSI in 256×256 patches with 50% overlap. Patches that contained no tissue, were classified as negative. If a patch contained nevus or melanoma classified pixels, it was run through the CNN. The CNN outputs a global probability for the slide belonging to each class. Since a slide can contain both melanoma and nevi structures, we determine the global label by placing higher importance on melanoma than nevi.

4. RESULTS

Performance of the U-Net was measured by comparing the model prediction to the manual ground truth annotations. Melanoma, nevus and negative probabilities were combined to a multi-class prediction map similar to the ground truth annotations. Some examples of the segmentation results are shown in Figure 1. Quantitative results were obtained by calculating the accuracy and Dice score between the prediction maps and ground truth within the manual annotation bounding box. The results are summarized in Table 2.

Table 2: Dice scores comparing segmentation results between the two scanners

Scanner	Metric	Melanoma	Nevus
Philips	Accuracy	0.92 ± 0.06	0.89 ± 0.08
	Dice	0.84 ± 0.08	0.81 ± 0.10
Hamamatsu	Accuracy	0.92 ± 0.06	0.90 ± 0.05
	Dice	0.84 ± 0.07	0.85 ± 0.09
Combined scanners	Accuracy	0.92 ± 0.06	0.89 ± 0.08
	Dice	0.84 ± 0.08	0.83 ± 0.10
Combined classes	Accuracy	0.915 ± 0.07	
	Dice	0.835 ± 0.09	

After calculating the segmentation probability maps on the validation set, we determine the global probability using the CNN. We then calculated the F1-score using a range of probability thresholds and we selected a threshold based on the highest score.

The CNN WSI classification was evaluated over the independent test set without manual segmentations. This set consists of 331 slides scanned on the Philips scanner for which only the diagnosis as either melanoma, nevus or negative from the clinical report was available. Using the threshold calculated over the validation set, we classified the WSI. The results are shown in Table 3. Out of 176 melanoma slides, 3 were misclassified as negative and out of 62 nevus 3 were classified as melanoma and 2 as negative.

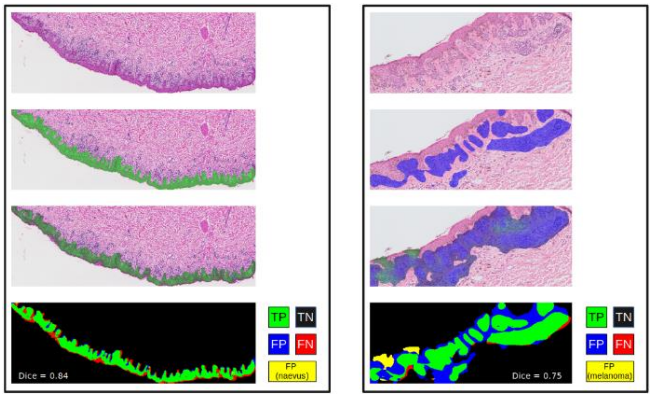

Figure 1: Example of segmentation results. On the left side segmentation results showing areas detected as melanoma. On the right side results for nevus detection. The model detects small structures on the epidermis as melanoma

Table 3: WSI classification on independent validation set

		Predicted		
		Melanoma	Nevus	Negative
Actual	Melanoma	173	0	3
	Nevus	3	57	2
	Negative	7	0	86

5. CONCLUSION & DISCUSSION

In this study, we developed and evaluated an automatic whole slide segmentation and classification system for skin excision and biopsy slides based on a combination of two deep models – a U-Net and a CNN. To the best of the authors' knowledge, this is the first system that achieves such high accuracy results on such a large dataset of WSIs[8], [11], [12].

Our model managed to not only accurately segment areas suspected for melanoma and nevus but also provide an accurate whole slide classification. We noticed however that it misclassified 3 slides out of 176. These 3 slides were segmented correctly by the U-Net but misclassified by the CNN. The reason is that these were early stage melanoma and quite small, something that our model apparently had not trained enough on. The nevi and negative slides that were misclassified were mainly due to tissue artifacts and inflammation detected by the model as melanoma areas. Even though we applied a hard negative mining step in our training process, the model had difficulty accurately segmenting those.

The segmentation provides a risk assessment overview and can also be used to measure Breslow thickness and margins although that was not explored in this study. The results are meant to assist a pathologist in either focusing on the areas with the highest risk within a slide, or, based on the

classification, to provide an overview of the risk associated with a case.

The study was limited as its clinical impact was not assessed and we did not explore the agreement of the model with a panel of specialists, only with the original diagnosis. A further limitation is that we only included clear-cut cases of either melanoma or nevus and it remains to be seen how well the model will perform in daily practice in which more challenging intermediate cases will be present, such as dysplastic nevus and Spitz tumors.

For future studies we will focus on assessing the clinical performance of the model with inclusion of more challenging cases as well as more types of skin malignancies most commonly encountered in daily practice.

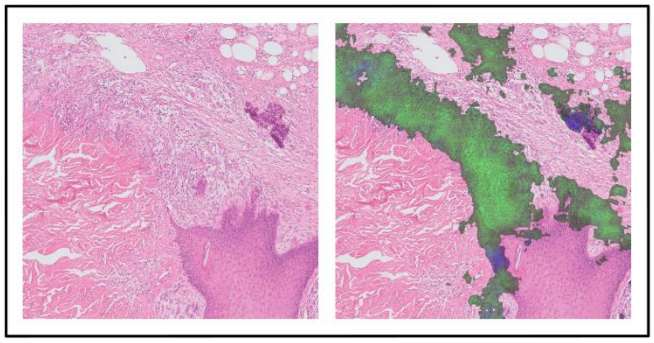

Figure 2: An example of a melanoma false positive on an area with inflammation. Green denotes melanoma detection probability and blue nevus.

6. REFERENCES

[1] IKNL, "Dutch cancer registration - Nederlandse Kankerregistratie." [Online]. Available: https://www.cijfersoverkanker.nl/nkr/index. [Accessed: 05-Sep-2019].

[2] R. Corona *et al.*, "Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions," *J. Clin. Oncol.*, vol. 14, no. 4, pp. 1218–1223, Apr. 1996.

[3] K. C. W. Veenhuizen, P. E. J. De Wit, W. J. Mooi, E. Scheffer, A. L. M. Verbeek, and D. J. Ruiter, "Quality assessment by expert opinion in melanoma pathology: Experience of the Pathology Panel of the Dutch Melanoma Working Party," *J. Pathol.*, vol. 182, no. 3, pp. 266–272, 1997.

[4] L. Brochez *et al.*, "Inter-observer variation in the histopathological diagnosis of clinically suspicious pigmented skin lesions," *J. Pathol.*, vol. 196, no. 4, pp. 459–466, Apr. 2002.

[5] B. A. Shoo, R. W. Sagebiel, and M. Kashani-Sabet, "Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center," *J. Am. Acad. Dermatol.*, vol. 62, no. 5, pp. 751–756, May 2010.

[6] M. C. R. F. Van Dijk *et al.*, "Expert review remains important in the histopathological diagnosis of cutaneous melanocytic lesions," *Histopathology*, vol. 52, no. 2, pp. 139–146, Dec. 2007.

[7] C. Andres *et al.*, "iDermatoPath – a novel software tool for mitosis detection in H&E-stained tissue sections of malignant melanoma," *J. Eur. Acad. Dermatology Venereol.*, vol. 31, no. 7, pp. 1137–1147, 2017.

[8] H. Xu, C. Lu, R. Berendt, N. Jha, and M. Mandal, "Automated analysis and classification of melanocytic tumor on skin whole slide images," *Comput. Med. Imaging Graph.*, vol. 66, pp. 124–134, 2018.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation BT - Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015," 2015, pp. 234–241.

[10] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 411–418.

[11] H. Xu, R. Berendt, N. Jha, and M. Mandal, "Automatic measurement of melanoma depth of invasion in skin histopathological images," *Micron*, vol. 97, pp. 56–67, Jun. 2017.

[12] A. Phillips and I. Teo, "Segmentation of Prognostic Tissue Structures in Cutaneous Melanoma using Whole Slide Images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, p. 0.