Deep Learning Segmentation of Invasive Melanoma

Aman Shah[a], Amal Mehta[a], Michael Wang[b], Neil Neumann[b], Avideh Zakhor[a], and Timothy McCalmont[b]

[a]University of California, Berkeley
[b]University of California, San Francisco

ABSTRACT

Melanoma is the deadliest skin cancer with the fastest rising incidence rate in the United States. The most important predictor of melanoma patient survival is the volume of invasive tumor at the initial biopsy. The appearance of in-situ melanoma in epidermis and invasive melanoma in dermis, which invades the underlying soft tissue and drives mortality, is often visually similar. We propose a novel two-stage method to segment invasive melanoma. The first stage computes two segmentation maps, one for tumor vs non-tumor and one for dermis vs epidermis. These two segmentation prediction maps of tumor and epidermis from the first stage combine to yield invasive melanoma predictions. Our method utilizes multiple resolutions and downscaling to increase information passed to the model and to improve model accuracy. Using an HRNet+OCR model for both epidermis and melanoma segmentation in our proposed two-stage system results in a marked improvement of F1 score (mIoU) to 0.44 (0.64) as compared to the current state-of-the-art of 0.14 (0.53).

Keywords: deep learning segmentation, invasive melanoma, epidermis, whole slide imaging, downscaling

1. INTRODUCTION

Melanoma is a lethal cancer with the fastest rising incidence of any cancer in the United States.[1] The national Surveillance, Epidemiology, and End Results cancer incidence data estimates 106,110 new cases in 2021.[2] Melanoma can invade through the first layer of the skin, called the epidermis, into the dermal layer below. Melanoma within the epidermis, called in-situ melanoma, is not known to cause death.[3] However melanoma in the dermis, called invasive melanoma (IM), could travel to distant body sites, in a fatal process called metastasis. For clarity, we use total melanoma (TM) to refer to the union of invasive and in-situ melanoma. The IM tumor burden is a crucial factor for patient survival prediction (staging) and clinical management.[4] Therefore, accurate assessment of IM tumor burden is paramount.

In current practice, IM tumor burden is assessed using Breslow thickness which is a measure of depth of tumor in the skin.[5] However Breslow et al.[5] reported that the incidence of metastasis is a function of the cross-sectional

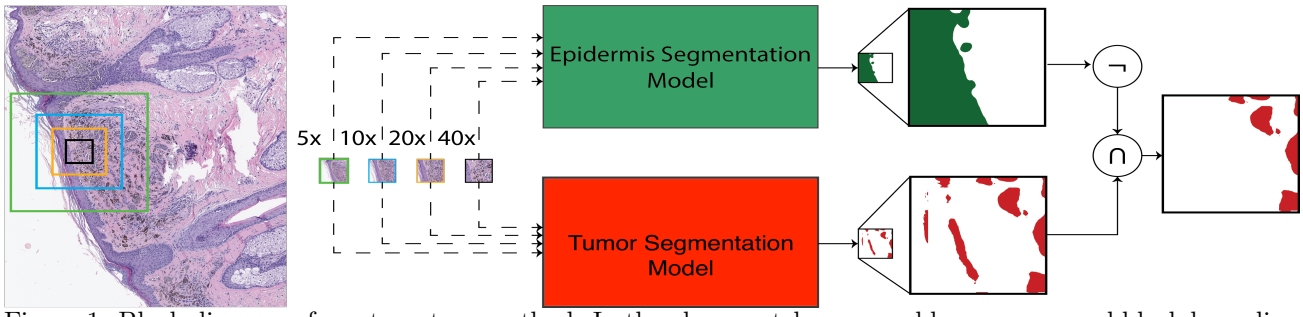


Figure 1: Block diagram of our two-stage method. In the above patches green, blue, orange, and black bounding boxes correspond to 5×, 10×, 20×, and 40× magnifications that are all downscaled to match the 40× patch size. Some subset of downscaled patches are passed into the models as indicated by the dashed lines. After the epidermis predictions are computed the outputs are negated, as indicated by ¬, and intersected with the outputs of the tumor segmentation model to create IM predictions.

tumor area, rather than a function of the thickness alone. Saldanha et al.[6] reported cross-sectional area to be an independent prognostic factor for survival.

Existing methods of estimating cross-sectional area are both time and labor intensive. Deep learning based segmentation offers an efficient method for evaluating the IM tumor cross-sectional area. This approach is also reproducible, which is necessary for practical adoption, and mitigates current ramifications stemming from high inter-observer variability from human attempts at segmentation.[7] In this paper we develop a two-stage method to segment IM, and verify our results from actual patients.

2. DATASET

This study has been approved by the institutional review board. We identified 57 University of California, San Francisco patients with IM for segmentation training and testing. In particular, these are primarily biopsies, stages T1 to T4, diagnosed between 2004 and 2014. These cases are single, individual melanomas and do not include melanomas co-existing with nevi or desmoplastic melanomas. Hematoxylin and eosin (H&E) stained slides of the specimens are digitized at $40\times$ magnification using a Leica Aperio whole slide scanner, with a final resolution of approximately 3960 pixels per millimeter, or a pixel size of $(15.68 \times 15.68)(\mu m)^2$. Identifiers associated with the pathology slides are removed at scanning. The whole slide images (WSIs) are annotated by two board-certified dermatopathologists using QuPath[8] to indicate regions of epidermis, IM, non-melanoma dermis (fibrosis, inflammation, normal dermis), and negative space (air, stratum corneum). The IM label includes mass forming tumors, infiltrative strands, and individual cells. Fine blood vessels and stroma, lymphocytes, and necrosis within the border of the tumor are labeled as IM. Skin appendages are labeled as epidermis.

3. METHOD

We propose a two-stage system to segment IM, as shown in Figure 1. The first stage consists of employing two semantic segmentation models, one for tumor vs non-tumor segmentation and one for epidermis vs dermis segmentation, respectively. Tumor vs non-tumor segmentation may refer to either IM vs non-IM or TM vs non-TM segmentation. The second stage intersects negative epidermis and positive tumor predictions, to create final IM predictions. This approach removes false positive IM predictions within the area denoted by the segmented epidermis region. As such it aims to locate IM in dermis which is fatal rather than in-situ melanoma in the epidermis which is harmless. As shown in the results section, this approach outperforms the classical one-stage method of directly segmenting for IM.

3.1 Data Augmentation

To make using models on large WSIs computationally tractable, we divide an input image into smaller, square patches.

Our training dataset for the task of epidermis vs dermis segmentation consists of 36 WSIs. Our training data exhibits an imbalanced pixel distribution of approximately 1:4 between the epidermis and negative class. To overcome this data imbalance in the training set, we alter the pixel distribution using sampling techniques.[9] Specifically, we remove "uninformative" patches containing less than 2% of cellular tissue. We further undersample patches corresponding to dermal regions to increase the proportion of epidermis. Lastly, we oversample patches with large areas of epidermis. In addition, we downscale patches and apply a combination of flips and rotations in multiples of 90 degrees which results in a pixel distribution of 3:7 between the epidermis and negative class. An additional 7 WSIs are separated into an independent epidermis testing set.

We use 43 WSIs for tumor segmentation training and apply two distinct approaches: (a) IM as the target class, and (b) with the union of IM and epidermis (IM $\cup$ E) as the target class as shown in Figure 1. IM $\cup$ E is a proxy for TM because such data is easier to annotate. Since we combine the melanoma predictions with the epidermis segmentation in the upper branch of Figure 1 either training scheme results in predictions of IM. We apply the aforementioned sampling and data augmentation techniques to both approaches resulting in a pixel distribution 4:6 between melanoma and the negative class. The remaining whole slide images are separated into an independent melanoma testing set.

After the H&E staining process WSIs may contain artifacts such as adhesive stains. Such artifacts are removed by applying Otsu's thresholding technique to the input image before training.[10]

Table 1: Stage 1, Epidermis Segmentation Results

#	Model	Input Patch Size	mIoU	fwIoU	F1 Score	Sensitivity	Specificity
1	3-Branched HookNet	800 × 800	0.76	**0.97**	0.69	0.91	0.98
2	HRNet+OCR	1576 × 1576	**0.79**	**0.97**	**0.74**	0.89	**0.99**
3	HRNet+OCR	3152 × 3152	0.77	**0.97**	0.72	0.95	0.98
4	UNet	1600 × 1600	0.75	0.96	0.69	**0.98**	0.98

Table 2: Stage 1, Invasive Melanoma Segmentation Results

#	Model	Input Patch Size	Targets	mIoU	fwIoU	F1 Score	Sensitivity
1	HRNet+OCR	788 × 788	Invasive Melanoma	0.5	0.94	0.1	0.49
2	2-Branched HookNet	800 × 800	Invasive Melanoma	0.51	**0.97**	0.08	0.06
3	3-Branched HookNet	800 × 800	Invasive Melanoma	**0.57**	**0.97**	**0.27**	0.36
4	UNet	788 × 788	Invasive Melanoma	0.43	0.80	0.08	**0.78**
5	Ref. 9	788 × 788	Invasive Melanoma	0.53	**0.97**	0.14	0.13

3.2 Network Architecture

In our experiments we utilize several architectures: UNet,[11] FCN variant designed by Ref. 9, two branched HookNet, three branched HookNet,[12] and HRNet+OCR.[13] HookNet[12] is a multi-resolution network with two encoder-decoder branches that take inputs of the same size but at different resolutions centered on a given pixel. We extend this work to allow for three resolutions by using three encoder-decoder branches. In our experiments, the two branched HookNet uses 40× and 10× magnification, while the three branched HookNet uses 40×, 10×, and 5× magnifications. Lastly, HRNet+OCR, a segmentation model that maintains high-resolution representations, was initialized with weights from a model pretrained on the CityScapes dataset.[14]

3.3 Training Details

We train our models using a weighted cross-entropy loss between ground truth and predictions. The weight for a given class is the inverse square root or the inverse of the class frequency. We use the Adam optimizer with learning rate 0.001, and beta parameters of 0.99 and 0.999 to update our weights throughout our experiments. We allocate 20 percent of the training data for a validation set.

Additionally, our HookNet models have multiple branches that each output a prediction, so we utilize a weighted sum of the cross-entropy loss from each branch. We used weights of 0.6, 0.4, and 0.3 for the 40×, 10×, and 5× branches, respectively. Our experiments using HookNet, UNet, and the FCN variant from Ref. 9 were trained for 150 epochs.

Our HRNet+OCR model also uses a weighted loss between two segmentation maps. HRNet+OCR outputs an additional set of predictions from an intermediate layer in the network, called the auxiliary layer. Incorporating predictions from the auxiliary layer into the loss improves the feature representation of the input, so we use weights of 0.4 and 0.6 for the auxiliary and final layer outputs of HRNet+OCR respectively.[13] Our experiments utilizing HRNet+OCR ran for 100 epochs.

4. EXPERIMENT RESULTS

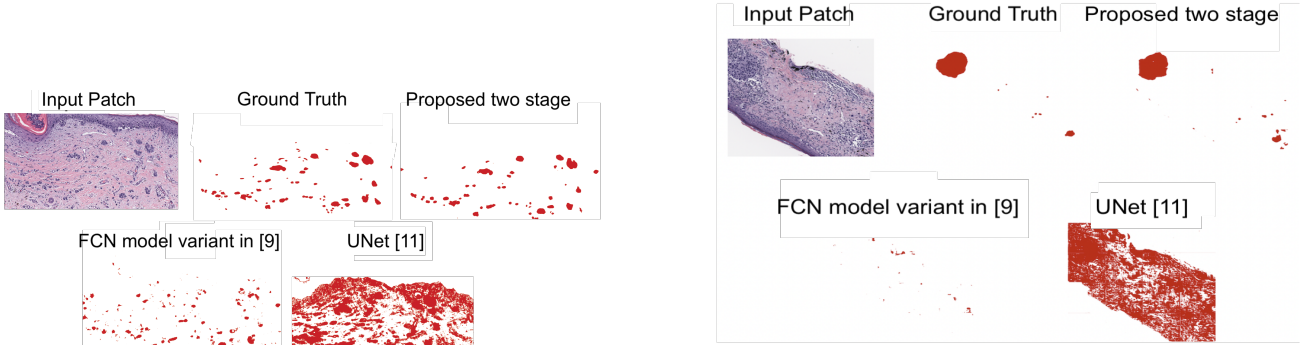
4.1 Epidermis Segmentation

Table 1 shows the results of our epidermis segmentation models, pictured in the upper branch of Figure 1 on our test set of 7 WSIs. The highest mIoU (F1 score) achieved is 0.79 (0.74), as shown in Table 1 row 2, resulting from the HRNet+OCR model with downscaling of the 1576×1576 patches to 788×788 patches. Furthermore, this epidermis model attains a sensitivity and specificity of 0.89 and 0.99 compared to 0.90 and 0.98 in Ref.9 which could make this a competitive model compared to other epidermis segmentation techniques.[15,16]

Although infrequent, presentations of epidermis such as epidermis surrounding hair follicles are circumstances where our model performs comparatively poor. This deficiency can be ameliorated by incorporating more adnexal structures in training.

Table 3: Stage 2, IM Segmentation Results; the number after the epidermis model architecture indicates the patch size which was downscaled before being passed as input to the model.

#	Stage 1 Epidermis Model	Stage 1 Tumor Model	mIoU	fwIoU	F1 Score	Sensitivity
1	HRNet+OCR 1576	HRNet+OCR (IM)	0.58	0.97	0.29	0.28
2	HRNet+OCR 1576	3-Branched HookNet (IM)	0.60	0.97	0.35	0.36
3	HRNet+OCR 1576	UNet (IM)	0.47	0.86	0.13	**0.74**
4	HRNet+OCR 1576	Ref. 9 (IM)	0.55	**0.98**	0.20	0.13
5	HRNet+OCR 1576	HRNet+OCR (IM ∪ E)	0.61	0.97	0.39	0.42
6	HRNet+OCR 1576	3-Branched HookNet (IM ∪ E)	0.58	0.97	0.32	0.50
7	HRNet+OCR 1576	UNet (IM ∪ E)	0.58	0.96	0.31	0.56
8	HRNet+OCR 3152	HRNet+OCR (IM)	0.56	0.97	0.24	0.22
9	HRNet+OCR 3152	3-Branched HookNet (IM)	0.62	**0.98**	0.41	0.37
10	HRNet+OCR 3152	UNet (IM)	0.49	0.89	0.15	0.18
11	HRNet+OCR 3152	Ref. 9 (IM)	0.56	**0.98**	0.22	0.14
12	HRNet+OCR 3152	HRNet+OCR (IM ∪ E)	**0.64**	**0.98**	**0.44**	0.39
13	HRNet+OCR 3152	3-Branched Hooknet (IM ∪ E)	0.62	0.97	0.41	0.37
14	HRNet+OCR 3152	UNet (IM ∪ E)	0.60	0.97	0.37	0.53
15	UNet	UNet (IM)	0.47	0.86	0.13	**0.74**
16	UNet	UNet (IM ∪ E)	0.57	0.97	0.27	0.29



(a) (b)

Figure 2: Two examples of IM predictions from the two-stage model from row 12 of Table 3, FCN model variant in Ref. 9 from row 5 in Table 2, and 2 stage UNet[11] from row 3 in Table 3 respectively.

4.2 Melanoma Segmentation

Table 2 shows the results of one-stage IM semantic segmentation models, pictured in the lower branch of Figure 1, on our test set of 13 WSIs. Results of our two-stage method, corresponding to the final output of Figure 1, are found in Table 3. The best result in terms of mIoU and F1 score among all the entries in Tables 2 and 3 is shown in row 12 of Table 3 and uses (a) the HRNet+OCR model which downscales patches from 3152×3152 for epidermis segmentation, in conjunction with (b) another HRNet+OCR model trained using IM ∪ E as the targets. This is noteworthy, indicating that the proposed two-stage approach has a 214% and 450% F1 score improvement over well known, traditional one-stage approaches such as Ref. 9 and UNet.

We speculate that HRNet+OCR performs best in our two-stage method because it uses a transformer-based network to generate strong high-level representations of the pixel-object relationship. The attention mechanism used by HRNet+OCR makes it superior to methods such as HookNet which use less sophisticated techniques such as nearest neighbor interpolation for upsampling. Furthermore, the benefit of information from larger neighborhoods that makes HookNet the best one-stage IM segmentation model, as shown in row 3 of Table 2, is offset by the epidermis model in our two-stage method which allows it to leverage information from larger patches and improved attention from HRNet+OCR.

Our best two-stage melanoma predictions in row 12 of Table 3 utilize a model trained with IM ∪ E as targets.

Comparing rows 1 vs 5, 3 vs 7, 8 vs 12, and 10 vs 14 in Table 3, we conclude that in the two-stage approach, keeping the epidermis model constant while changing the tumor model targets from IM to IM $\cup$ E leads to improved mIoU and F1 scores. We speculate this difference is caused by in-situ melanoma. Specifically, treating IM as the target class degrades training since visually similar IM and in-situ melanoma are classified differently. When using TM, or our proxy IM $\cup$ E, as targets, IM and in-situ melanoma are the same class which utilizes the similarity between melanomas. Using larger patches or downscaling to address the similarity of these melanoma is not feasible due to small clusters of melanoma and computational constraints.

Comparing entries in Table 2 and the top 4 rows in Table 3, we note that using an IM segmentation network in our two-stage method can always improve mIoU as compared to the corresponding one-stage IM segmentation. This shows that our two-stage approach can be used to improve one-stage segmentation performance in an architecture independent way. These improvements are caused by the reduction of false positives that occur due to in-situ melanoma.

As seen in rows 1-7 and 8-14 of Table 3, patches downscaled from 3152×3152 in our two-stage method yields higher mIoUs, fwIoUs, and F1 scores compared to 1576×1576 patches in 6 out of 7 cases. This indicates that information from a larger neighborhood is beneficial for the epidermis model.

Our system can successfully identify small clusters or single cells of IM, as seen in Figure 2. However, a current limitation is detection of small melanoma clusters obfuscated by high background inflammation or fibrosis.

5. FUTURE WORK

The next steps include testing this cross section area hypothesis as a predictor of cancer stage with 80% power at 5% significance level, which will require 197 test samples assuming the standard deviation of the IoU is 0.5. The deep learning method presented and manually measured Breslow thickness will be compared. The deep learning calculated cross-sectional area of IM developed in this paper will be used to develop a survival prediction. We will extend this study to create a tool to assist medical professionals with end-to-end processing and analysis of medical data. This tool could improve prognostic prediction and management guidance for patients, as well as potentially revise the T-category staging system.

REFERENCES

[1] International Agency for Research on Cancer, "GLOBOCAN 2012 v1.0, cancer incidence and mortality worldwide: IARC CancerBase No. 11.." http://globocan.iarc.fr. "(Accessed: 2 June 2021)".

[2] National Cancer Institute, "Cancer Stat Facts: Melanoma of the Skin." https://seer.cancer.gov/statfacts/html/melan.html. "(Accessed: 19 May 2021)".

[3] Giblin, A.-V. and Thomas, J., "Incidence, mortality and survival in cutaneous melanoma," *Journal of Plastic, Reconstructive & Aesthetic Surgery* **60**(1), 32–40 (2007).

[4] Mccalmont, T. H., "The second dimension—integrating calculated tumor area into cancer diagnosis," *JAMA Dermatology* **155**(8), 883 (2019).

[5] Breslow, A., "Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma," *Annals of Surgery* **172**(5), 902–908 (1970).

[6] Saldanha, G., Yarrow, J., Elsheikh, S., O'Riordan, M., Uraiby, H., and Bamford, M., "Development and initial validation of calculated tumor area as a prognostic tool in cutaneous malignant melanoma," *JAMA Dermatology* **155**(8), 890–898 (2019).

[7] Joskowicz, L., Cohen, D., Caplan, N., and Sosna, J., "Inter-observer variability of manual contour delineation of structures in ct," *European Radiology* **29**, 1391–1399 (2018).

[8] Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., Mcart, D. G., Dunne, P. D., Mcquaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., and et al., "Qupath: Open source software for digital pathology image analysis," *Scientific Reports* **7** (Dec 2017).

[9] Phillips, A., Teo, I., and Lang, J., "Segmentation of prognostic tissue structures in cutaneous melanoma using whole slide images," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*], (June 2019).

[10] Otsu, N., "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1), 62–66 (1979).

[11] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*], Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., eds., 234–241, Springer International Publishing, Cham (2015).

[12] van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., and Ciompi, F., "Hooknet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Medical Image Analysis* **68**, 101890 (2021).

[13] Yuan, Y., Chen, X., and Wang, J., "Object-contextual representations for semantic segmentation," in [*16th European Conference Computer Vision (ECCV 2020)*], (August 2020).

[14] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B., "The cityscapes dataset for semantic urban scene understanding," in [*Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (2016).

[15] Xu, H., Berendt, R., Jha, N., and Mandal, M., "Automatic measurement of melanoma depth of invasion in skin histopathological images," *Micron* **97**, 56–67 (March 2017).

[16] Amor, R. D., Morales, S., Colomer, A., Mogensen, M., Jensen, M., Israelsen, N. M., Bang, O., and Naranjo, V., "Automatic segmentation of epidermis and hair follicles in optical coherence tomography images of normal skin by convolutional neural networks," *Frontiers in Medicine* **7** (2020).