

This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Segmentation of Prognostic Tissue Structures in Cutaneous Melanoma using Whole Slide Images

Adon Phillips EECS aphil037@uottawa.ca Iris Teo Pathology and Lab. Medicine iteo@eorla.ca Jochen Lang EECS jlang@uottawa.ca

University of Ottawa, Canada

Abstract

Our work applies modern machine learning techniques to melanoma diagnostics. First, we curated a new dataset of 50 patient cases of cutaneous melanoma in whole slide images (WSIs). We applied gold standard annotations for three tissue types (tumour, epidermis, and dermis) which are important for the prognostic measurements known as Breslow thickness and Clark level. Then, we devised a novel multi-stride fully convolutional network (FCN) architecture that outperformed other networks trained and tested using the same data and evaluated on standard metrics. Three pathologists measured the Breslow thickness on the network's output. Their responses were diagnostically equivalent to the ground truth measurements, showing that it is possible to overcome the discriminative challenges of the skin and tumour anatomy for segmentation. Though more work is required to improve the network's performance on dermis segmentation, we have shown it is possible to achieve a level of accuracy required to manually perform the Breslow thickness measurement.

1. Introduction

Cutaneous melanoma (CM, melanoma) is an aggressive form of skin cancer originating in melanocytes which are the cells responsible for the pigmentation of skin, hair and eyes. Melanoma develops as the result of DNA damage usually as a consequence of intense exposure to ultraviolet radiation. When a melanocyte suffers sufficient genetic damage it may begin to grow in an abnormal way and become a melanocytic tumour [16]. In the earliest stage of development the tumour will grow laterally along the outer epidermal layer of the skin. At this point the complete surgical excision of the lesion is possible. However, if left undiagnosed the tumour may thicken and infiltrate deeper into the skin. If the tumour growth invades lymphatic vessels, cancer cells may separate the matrix the primary tumour and travel to regional lymph nodes as metastases in-transit. The tumour may also invade blood vessels causing cancer cells to be transported to distant locations resulting in metastatic cancer [3].

In the United States melanoma is the second most common cancer in men aged 20-39 and the fifth most common in men of any age. In women it is the third most common cancer through ages 20-39 and the sixth most common in women of all ages [4, 1]. In places such as New Zealand and Australia, melanoma incidence is reaching epidemic levels especially in young people [11]. Fortunately, if diagnosed early, the five year survival rate is as high as 98%. If the tumour has spread to regional nodes, five year survival is reduced to 62% and then to only 18% in the case of metastatic disease [4]. While melanoma represents 3% of all skin cancers, it is responsible for 65% of skin cancer deaths [27]. Clearly the importance of early detection and accurate diagnosis cannot be overstated.

The gold standard for the diagnosis of melanoma requires full-thickness excisional biopsy of suspected lesions followed by the histopathologic analysis of hematoxylin and eosin (H&E) stained tissue sections [12]. Historically, and to this day this analysis is conducted under a light microscope for the vast majority of cases. However the recent availability of slide scanning technology, which is able to digitize prepared tissue specimens on glass slides into high resolution files has brought the medical specialty of pathology into the midst of a transition towards a digital workflow. As a consequence, this move to 'Digital Pathology' (DP) has also afforded the computational image analysis community access to whole slide images (WSI) allowing the application of modern image processing techniques to the processing of such specimens [22]. This includes automated, computer-aided, and other diagnostic tools that have the ability to augment current clinical workflows in the hopes of improving patient care and outcomes.

A major component of the diagnostic process is a pattern recognition task where the pathologist uses visual infor-

mation combined with deep domain knowledge to identify complex histologic and cytologic features. In the case of cancer diagnosis, the variance in slide preparation, nuances of anatomy relating to qualitative and quantitative criteria of diagnosis, not to mention variability introduced by different slide scanners and associated software processing creates a feature space that is beyond the capacity of most handcrafted descriptors/extractors to account for. It has been shown that deep learning techniques have the potential to overcome such challenges, but in the case of melanoma diagnosis in WSIs which to our knowledge has not been explored, we will demonstrate that it is possible to accurately predict segmentation masks for multiple tissue types whose structures are vastly different in scale, and morphology with dramatic variations even between tissues of the same type as is the case in the pathology of the skin (see Figure 1).

Furthermore, in addition to the diagnosis there is also a list of pertinent features for the prognostication and management of a given disease. In the case of melanoma, the most important prognostic factor is the maximum tumour thickness, sometimes referred to as the Breslow thickness. According to the College of American Pathologists melanoma reporting protocol, considered throughout North America as the standard for reporting [10]. While we do not know the segmentation accuracy required for such measurements in prognosis, we will show that it is possible to produce segmentation masks that are qualitatively, of sufficient accuracy to perform the Breslow thickness measurement.

2. Related Work

Much of the work in digital pathology involves the detection/segmentation of cellular structures such as nuclei, mitoses, individual cells and in our case epidermis/dermis layers of the skin. The motivation for identifying these structures relates to their role in various quantification methods as part of the diagnosis or staging of disease. Fully Convolutional Networks (FCNs) have excelled in such scenarios performing feature detection for segmentation in medical images [33, 24, 19]. Architectures such as UNet [33] and FCN-16s [21] that use skip connections have improved the quality of the output segmentations. Techniques such as using multiple scales [7, 18, 29] incorporate additional training information. While other approaches such as deconvolutional networks [26, 6] improve segmentation results by learning the deconvolutional step used to recover pixel resolution information lost during convolution and pooling stages. Our approach incorporates deconvolution at multiple levels of granularity to produce the final segmentation mask. This results in improved segmentation at course and fine structures while retaining the efficiency of the FCN network by Long *et al*.

There has been limited application of these techniques to melanoma in whole slide images, though in the case of epidermis segmentation there has been some work done using non-convolutional methods. Xua *et al.* [31] perform epidermis segmentation using a thresholding, and k-means approach. The authors extracted features from epidermis and dermis tiles to classify melanocytic tumours using an SVM model. Kłeczek *et al.* [14] developed a shape-based technique to segment epidermis. They improved the base segmentation by analyzing the H&E stain concentrations on the slides. They report mean sensitivity at 87%, specificity at 95%, and precision at 57% on their test dataset for epidermis tissue. We show that using a convolutional network can provide accurate epidermis segmentation masks.

3. Dataset

We have assembled a dataset that includes 50 whole slide images (WSIs) containing H&E stained, full-thickness excisional biopsies from 49 individuals. Source images were scanned at $40 \times$ objective magnification using Aperio ScanScope slide scanners from nine different medical institutions. Annotations are provided at the equivalent resolution to the $40 \times$ magnification source images.

The unlabelled source images for the dataset were obtained from the The Cancer Genome Atlas (TCGA). The images were selected from the over one thousand cutaneous melanoma slides from the human skin cutaneous melanoma (HSCM) TCGA sub-project. Due to the wide variation in slide quality available from the HSCM project we selected a smaller set of 150 candidate slides that were evaluated for inclusion in our dataset based on the exclusion of material with frozen section artifact (the dataset does not include any tissues from frozen sections), and exclusion of slides with extensive tissue folding or other artifacts that made the histology uninterpretable. From these 150 candidates, 50 WSIs were included in the final dataset based on the slide containing each of the target tissue types. The dataset was limited to 50 WSIs due to the time and effort required from an expert for the annotation step.

The overall dataset is divided into training, validation, and test sets at a ratio of 70:15:15 percent respectively. This results in 36 WSIs for training, and 7 WSIs each for validation and testing. The data set was stratified by patient across each dataset. Due to the data size and gigapixel resolution of the WSIs, each slide is tiled to create image patches at 512×512 pixel resolution to facilitate GPU processing. Patches are non-overlapping palette-based png files. Each patch required post-processing to ensure the five colour palette indices were in the same order. Patches that did not contain tissue matter were discarded from the training set. In Table 1 we provide basic image and patch statistics for the test set.



Figure 1: A selection of patches to illustrate the variation between structures and staining for each tissue type in our data set.

Case ID	Size (MB)	Pixel Size	# of Patches
A5EO	536.5	15438×11582	660
A3Z4	1146.1	23086×16546	1440
A3XE	1098.1	21420×17087	1353
A2GO	1407.4	24751×18952	1776
A182	1761.8	27845×21088	2214
A19H	2276.1	30463×24903	2832
A8K4	1937.4	29381×21978	2394

Table 1: Image details for our WSI test cases.

3.1. Annotations

Colour mask, gold standard annotations were applied over the source images. The annotations were manually defined using three colors in the RGB colour space representing three tissue classes. Red (#FF0000) denoting tumour regions, yellow (#FFFF00) denoting the dermis layer, and green (#00FF00) denoting the epidermis. The overall annotation also includes two additional colours to indicate a background class and a "not diagnostically important" (NDI) class. These are denoted by white (#FFFFFF) and black (#000000) respectively. The NDI regions include all tissue areas (including artifacts on the slide e.g. pen markings) not marked by the pathologist (note that the stratum corneum is considered NDI). The background class represents only regions without tissue present.

As most of the selected WSIs contained tumours that were mass forming, the external edge of the tumour was outlined. Large areas which were composed of inflammatory cells were not marked as tumour; however, if melanoma cells resided within the inflammatory population and could be positively identified, these individual cells were marked. The fine blood vessels and stroma of the tumour were also included with the tumour, as were tumour infiltrating lymphocytes within the borders of the tumour. Necrotic areas, apoptotic cells, and any areas of cystic degeneration were also marked as tumour. At ulcer sites, only tumour cells were marked, and areas of fibrinopurulent exudate were not marked as tumour. Where they could be positively distinguished from mimickers, melanoma cells within the epidermis (in-situ melanoma) were also marked but were not actively sought out.

The epidermis was labeled from the basement membrane to the top of the granular layer (the outer edge of the epidermis). Markedly reactive epidermis and tangentially sectioned epidermis was included. The stratum corneum (the outermost layer of the skin) was not labelled. As the infundibular portion of the hair follicle derives from the epidermis, this was included in the epidermis labels. In spite of its morphologic similarity to the epidermis, the remainder of the hair follicle was included within the dermis layer as it represents a dermal structure. This decision requires further analysis to determine the impact on the prediction of the included tissue types.

While the dermis was normal in many cases, in others, scar or alteration by tumour was found. All of these were eventually included in the definition of the dermal layer. The lower boundary of the dermal layer was also difficult to define, as it undulates and merges imperceptibly with the fibrous septae of the underlying subcutaneous adipose tissue and an arbitrary cutoff had to occasionally be employed.

3.2. Data Augmentation

The thirty six patient cases included in the training set translates to over 62,000 image patches or approximately 16.3 billion total pixels. Based on other similar works using semantic segmentation [13, 8, 17, 23], this represents a sufficiently large collection of data for experimentation. However, as is the case with many medical imaging datasets, due to the natural distribution of the tissue types in our dataset, there exists a class imbalance between tissue types. Our initial dataset suffered from two imbalance problems. First, since WSIs consist of a tissue specimen placed in the cen-

ter of the slide, and due to the background being visible through the tissue specimen, background pixels are the most common by far. Background pixels outnumber the second most common class (tumour) 4:1. Most of these pixels are in patches that only contain background pixels. Secondly tumour labelled pixels outnumber epidermis labelled pixels 28:1 (see Figure 2). Initial experiments with the unmodified data showed a test network would converge with high accuracy by predicting the tumour class for all tissue types (severe overfitting). Furthermore, early experiments also showed there were insufficient pixels labelled in the epidermis class to train the network to recognize such regions accurately in the face of such a large imbalance.

To address these problems we used a class balancing solution to effectively undersample and oversample the problematic classes as required:

We undersample by removing patches where background pixels are most common and make up more than 98% of the target patch. This percentage retains patches that might contain locally important features. We also remove patches where background pixels are most common and the NDI class is the only other class present.

For the epidermis class, which is the minority class, we oversample using basic data augmentation transformations to increase the number of epidermis containing patches with respect to patches of other classes. We parse the pixels of each patch containing the annotations and store the list of patches that meet the following criteria:

- All pixels are epidermis
- Most pixels are background, second most pixels are epidermis
- Most pixels are background, second most pixels are NDI, third most pixels are epidermis
- Most pixels are NDI, second most pixels are epidermis
- Most pixels are NDI, second most pixels are NDI, third most pixels are epidermis

These annotation patches and the corresponding feature patches are then augmented with the following transformations: Flip left to right, flip top to bottom, rotate 90° , rotate 270° .

The result of the proposed balancing scheme is a generally balanced dataset with a modest but tractable imbalance remaining between the tumour and epidermis classes. This imbalance was reduced from 28:1 to 7.75:1 (see Figure 2). It might seem to make sense to undersample the tumour class to further reduce the imbalance, though after experimentation we found that the epidermis was receiving accurate segmentations compared to pre-balancing tests. Furthermore the anatomical variance of tumour cells and regions across patients and even individual slides is significantly greater than that of the epidermal cells (see Figure 1 for an illustration of in-class variation). Therefore it was advantageous to retain as many tumour examples as possible.



Figure 2: Class distribution by the number of pixels per class following the described data augmentation strategy.

4. Network



Figure 3: A graphical overview of multiple deconvolutional blocks linked together. Scores from each deconvolution are combined using element-wise summation. This configuration represents that used to train our model. Skip connections shown originate at the third and fourth pooling layers (left to right) in the standard FCN network from Long *et al.* [21].

To perform the segmentation we used a modified version of the fully convolutional network design by Long *et al.* [21]. Using the FCN style network of Long *et al.*, three levels of output granularity can be obtained based on three architectural variants in three discrete networks. These three network configurations range from a single stream with a 32 pixel stride (FCN-32s) to three streams fused into an 8 pixel stride (FCN-8s). In FCN-16s, scores from each stream are fused in a step-wise fashion so that intermediate scores from stream 1 and stream 2 are combined and upsampled using deconvolution, then that result is fused before being upsampled with a stride of 16 resulting in the final output score. While Long et al.'s FCN-16s includes information from multiple streams the final output granularity is fixed. To incorporate information at multiple levels of prediction granularity we compute three levels using transposed convolution (also known as deconvolution) at three independent pixels strides (32, 16, 8). The result is segmentation information for features of various scales, that in our testing reduces errors compared to using a single level of granularity to capture features at all scales. For example, using FCN-8s (most granular) alone results in noisy output for large contiguous regions of a single class, and FCN-32s (most coarse) can't capture fine details in our dataset. To make use of this information we combine these results using a weighted, element-wise summation that combines the output maps. The contribution of each granularity is controlled via per-stride weights as hyper-parameters for our model (see Figure 3). This approach out-performs the standard FCN-16s design on this task.

4.1. Transfer learning

We use the PascalVOC segmentation dataset [9] to pretrain our network. Notwithstanding the significant difference in domain specific features between the PascalVOC dataset and our medical images, transfer learning was essential to improving our training times and results. We made extensive use of transfer learning to reduce training time and improve results. We initialized weights for learnable layers using publicly available models (FCN-32s, FCN-16s, FCN-8s) and fine-tuned from there. For our multi-stride networks we initialized weights using multiple pre-trained networks trained at each requisite stride for the given branch of our network. Non-matching weights were learned from random initializations.

4.2. Evaluation Metrics

In order to evaluate the performance of our proposed architecture on our curated dataset we have selected a set of commonly used metrics [30], designed to measure the performance of semantic segmentation inference results. We compute: Pixel accuracy, Mean Pixel Accuracy (mPA) as:

$$\frac{1}{n_{cl}} \frac{\sum_i n_{ii}}{t_i}$$

Mean Intersection Over Union (mIoU) as:

$$\frac{1}{n_{cl}}\sum_{i}\frac{n_{ii}}{(t_i+\sum_j n_{ji}-n_{ii})}$$

and Frequency Weighted Intersection Over Union (fwIoU) as:

$$rac{1}{\sum_k t_k} rac{\sum_i t_i n_{ii}}{(t_i + \sum_j n_{ji} - n_{ii})}$$

where n_{cl} represents he number of classes, t_i ; the total number of pixels in class i, and n_{ij} ; the number of pixels of class i predicted to belong to class j. So for class i and j: let n_{ii} represent the number of correctly classified pixels (TP) precision, n_{ij} ; the number of pixels wrongly classified (FP), and n_{ji} : The number of pixels wrongly not classified (FN). Finally, inspired by Zhou *et al.* [34], we also compute a 'Final Score' as the definitive metric for our evaluation as: $\frac{mPA+mIOU}{2}$

In addition to these metrics we also compute the area under the receiver operator curve (AUC) [5] for each class, and we show the confusion matrix for the test set.

5. Results and Discussion

5.1. Model Accuracy

Using the described pipeline we processed each WSI from our test set. We assembled the inference results into a whole slide prediction mask and computed metrics to assess the accuracy and the quality of the model. The test set was selected randomly and represents WSIs from multiple different slide scanners, and associated slide preparation processes.

Considering the results in Table 2, and Figure 5 we observe that the model is consistently able to provide a high quality for the whole slide. The variation in performance across slides could be explained by the large variability in the processes that produced the training images obtained from the TCGA. This could be viewed as a closer representation of what one might expect in a real-world environment. As such, using the TCGA data could be considered a strength of this approach, compared to the practice of preparing a dataset in-house using consistent hardware and associated physical processes. Increasing the size of the overall dataset could assist to improve the model in this regard.

5.2. Tissue Accuracy

Based on per-tissue results summarized in Figure 4 and Table 3 the model is able to differentiate between tissue types. There is a strong correspondence between the gold standard, and the model's predictions.

In the case of tumour regions the model is able to localize each structure and find accurate slide level boundaries. However, areas around patch boundary regions are prone to error though this does not affect practical accuracy for our task, but can lead to error accumulation. Tumour masses are accurately detected with most false positive regions consisting of small clusters that could be removed in a post processing step.

The model performs very well on epidermis segmentation. The model is able to find the contours of rete ridges of the epidermis (finger-like structures of the epidermis), and avoid the stratum corneum layer which was not annotated as epidermis Figure 7. The method could function as a standalone approach for epidermis segmentation as shown in the work of Xu and Mandal [32] and Kłeczek *et al.* [14]. For comparison our model has a sensitivity of 90%, specificity of 98%, and precision of 64% on the task of epidermis segmentation alone.

In the case of dermis tissue, the model is able to clearly define the superficial boundary, but the deep boundary is poorly differentiated. It should be noted that while the epidermis serves as the superficial boundary of the dermis, the lower (deep) boundary of the dermis is traditionally the adipose tissue. This is challenging to annotate, as the dermal collagen merges with the fibrous septae of the subcutaneous adipose tissue resulting in an arbitrary distinction as to the deep boundary of the dermis. More work is needed to devise a strategy for the annotation of this area in the future (see Figure 6 for a visualization of where the errors appear in the dermis).



Figure 4: The confusion matrix based on the predictive performance on the testset. The matrix shows a strong corelation between the gold standard annotations and the model's predictions. The NDI class is omitted from the matrix.

Case ID	PA	mPA	mIoU	fwIoU	Score	Time (s)
BF-A5EO	0.9640	0.6887	0.6454	0.9479	0.6670	48.4411
D9-A3Z4	0.9119	0.5976	0.5516	0.8635	0.5746	105.7572
EB-A3XE	0.9221	0.6165	0.5715	0.8810	0.5940	100.3250
EE-A2GO	0.9148	0.6176	0.5749	0.8844	0.5963	131.6688
EE-A182	0.8724	0.6180	0.5563	0.8598	0.5872	164.6194
ER-A19H	0.8775	0.7056	0.6496	0.8435	0.6776	210.1503
WE-A8K4	0.7742	0.5243	0.4975	0.7638	0.5109	177.4948
Mean:	0.8910	0.6240	0.5781	0.8634	0.6011	134.0652

Table 2: Calculation of our selected standard metrics for inference results. The results were obtained from the test set consisting of seven whole slide images of various sizes and class composition. See subsection 4.2 for a description of the metrics. Time is measured from the point each patch enters the network until the output is calculated. The final time is the sum of this measure for all patches of a given WSI from the test set.

Case ID	Epidermis	Dermis	Tumour	Background
BF-A5EO	0.9022	0.7811	0.9698	0.9890
D9-A3Z4	0.9407	0.9188	0.9724	0.9357
EB-A3XE	0.9562	0.9164	0.9798	0.9328
EE-A2GO	0.9702	0.8611	0.9826	0.9599
EE-A182	0.9230	0.7058	0.9138	0.9849
ER-A19H	0.9811	0.8191	0.9639	0.9601
WE-A8K4	0.9234	0.6490	0.7435	0.9838

Table 3: AUC values for each WSI in the testset.

5.3. Breslow Measurement

The primary goal of this work was to provide segmentations for prognostic tissue structures. To assess the accuracy of the model's predictions for this task we a assembled a panel of four pathologists to measure the Breslow thickness on the output and the results were diagnostically equivalent. When removing cases where measurements were not appropriate, inter-rater agreement was 87.5% with a fixed marginal kappa [28] of 0.5. Following the assessment a consensus-based assessment resulted in 100.0% agreement between pathologists that the true breslow thickness for each slide could be measured using the model's predictions when the measurement is performed correctly. This result supports the claim that the tumour and epidermis segmentation masks are sufficiently accurate to manually measure the Breslow thickness and thus it would be possible to use our approach as the basis for machine-based measurement of the Breslow thickness as part of an automated diagnostic workflow.

5.4. Network Comparison

To evaluate our network design we compared the results against the baseline FCN-16, and an FCN conversion of AlexNet [15]. While these networks no longer represent the state-of-the-art, they are commonly used in medical imaging due to their relatively low computation requirements and ease of configuration [20, 25]. All networks were



Figure 5: Demonstration of the multi-stride network using four whole slide images from the testset.



Figure 6: Visualization of segmentation errors by pixel. White represents true positive, teal represents false positive and red represents false negative by respective tissue type. The figure highlights the model's difficulty finding the boundary between the dermis and deeper tissues within the skin.

trained and tested using the same datasets, and were initialized with their respective publicly available pre-trained weights based on PascalVOC.

Our network showed superior performance to the comparators, based on our selected standard metrics (see Table 4). The performance improvement derives from the multi-strided upsampling information that is better able to capture a mix of both coarse and fine details, even difficult contextual situations (see Figure 8).

6. Conclusions and Future Work

We have demonstrated a method capable of detecting and localizing melanoma tumours and epidermis structures important for the Breslow thickness prognostic measurement using a custom convolutional neural network on a dataset that we have curated.

Given the qualitative and quantitative results, we show that it is possible to overcome the discriminative challenges of the skin and tumour anatomy for segmentation. Further



Figure 7: Examlple of the model's segmentation performance (right) on epidermis tissue. The black spot is an artefact present on the slide.

Network	PA	mPA	mIoU	fwIoU	Score	Time (s)
AlexNet	0.8904	0.5300	0.4913	0.8495	0.5106	51.3423
FCN-16s Our Multi-Stride	0.8775 0.8910	0.5513 0.6240	0.5068 0.5781	0.8433 0.8634	0.5291 0.6011	139.5002

Table 4: A summary of the network inference results showing the mean values computed across each of the test slides for each respective network. Bold text represents the best result for each metric. The final time is the sum of this measure for patches from a given WSI in the test set, and only includes the time to run the forward pass through the network at inference (*i.e.* not file IO). We note the inference time is slightly faster for our model. This suggests that our multi-stride fusion network was able to learn a less complex function (filter) mapping than the FCN-16s network.



Figure 8: Example of segmentation differences from a randomly selected patch. Note that our network is able to identify the small dermis (yellow) regions embedded in the epidermis, while the other networks were unable to correctly segment these areas.

we have shown it is possible to approach a level of accuracy required to allow manual measurement of the Breslow thickness as a precursor to automation. More work is required to improve the network's performance on dermis segmentation.

6.1. Future work

While we look forward to a direct augmentation of the human diagnostic process using works similar to that presented herein, the application to quality assurance and control (QA/QC) within a pathology laboratory represents a great opportunity. While the details of QA/QC processes vary widely between institutions, most quality assurance schemes utilize either prospective or retrospective review of past cases for diagnostic discrepancies. This requires both recognition and investment of medical personnel from medical laboratories to provide the resources required for such a program.

We foresee a role for this work and those of its kind to act in a pre-screening or post-screening capacity, potentially to flag or screen cases for problematic aspects and prioritize or facilitate for human review. Working with patient cases in a QA/QC context is a logical application area since there is no direct connection to diagnosis and treatment, and therefore no direct impact on patient safety in the case of a failure. While this work is not sufficiently developed to automate such aspects of QA/QC we feel this is a likely direction for this work as a precursor to inclusion in a clinical diagnostic workflow.

7. Acknowledgements

This research was partly funded by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

- [1] Melanoma stats, facts, and figures. *Learning and Instruction Symposium 2005, Joensuu, Finland.* 1
- [2] 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016.
- [3] M. D. Abeloff, J. O. Armitage, J. E. Niederhuber, M. B. Kastan, and W. G. McKenna. Ch 73. Melanoma. In *Abeloff's Clinical Oncology*, pages 1229–1252. Churchill Livingstone Elsevier, Philadelphia, PA, 2008. 1
- [4] American Cancer Society. Cancer facts & figures 2016. 04/04/2017. 1
- [5] D. Ballabio, F. Grisoni, and R. Todeschini. Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, 174:33–44, 2018.
 5
- [6] H. Chen, X. Qi, L. Yu, and P. Heng. DCAN: deep contouraware networks for accurate gland segmentation. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016 [2], pages 2487–2496. 2
- [7] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016 [2], pages 3640–3649.

- [8] A. Cruz-Roa, A. Basavanhally, F. Gonzalez, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *SPIE Medical Imaging*, 9041:904103–904103, 2014.
- [9] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303– 338, 2010. 5
- [10] D. Frishberg, C. Balch, B. Balzer, A. Crowson, M. Didolkar, J. McNiff, R. Perry, V. Prieto, P. Rao, M. Smith, B. Smoller, and M. Wick. Protocol for the examination of specimens from patients with melanoma of the skin. *Archives of Pathol*ogy and Laboratory Medicine, 133(10):1560–1567, 2009. 2
- [11] A. Glazer, R. Winkelmann, and A. Farberg.... Analysis of trends in US melanoma incidence and mortality. *Journal of the American Medical Association dermatology*, 153(2):225–225, 2017. 1
- [12] A. Goodson and D. Grossman. Strategies for early melanoma detection: approaches to the patient with nevi. *Journal of the American Academy of Dermatology*, 60(5):719–719, 2009. 1
- [13] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7:29, 2016. 3
- [14] P. Kleczek, G. Dyduch, J. Jaworek-Korjakowska, and R. Tadeusiewicz. Automated epidermis segmentation in histopathological images of human skin stained with hematoxylin and eosin. In M. N. Gurcan and J. E. Tomaszewski, editors, *Medical Imaging 2017: Digital Pathology, Orlando, Florida, United States, 11-16 February 2017*, volume 10140 of *SPIE Proceedings*, page 101400M. SPIE, 2017. 2, 6
- [15] A. Krizhevsky, I. Sutskever, and G. Hinton. *ImageNet classification with deep convolutional neural networks*. Curran Associates Inc., 2012. 6
- [16] V. Kumar, A. Abbas, N. Fausto, and J. Aster. *Robbins and Cotran: Pathologic Basis of Disease*. Elsevier Saunders, 2014. 1
- [17] L. Liang, G. Wang, R. Zhang, R. Zhang, X. Liang, and W. Zuo. Deep structured scene parsing by learning with image descriptions. *CoRR*, abs/1604.02271, 2016. 3
- [18] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5168–5177, July 2017. 2
- [19] G. Litjens, C. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, 6:26286, 2016. 2
- [20] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. 6
- [21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 3431–3431, 2015. 2, 4

- [22] A. Madabhushi and G. Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170 – 175, 2016. 1
- [23] B. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M. Weber, T. Arbel, B. Avants, N. Avache, P. Buendia, D. Collins, N. Cordier, J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. Iftekharuddin, R. Jena, N. John, E. Konukoglu, D. Lashkari, J. Mariz, R. Meier, S. Pereira, D. Precup, S. Price, T. Raviv, S. Reza, M. Ryan, D. Sarikava, L. Schwartz, H. Shin, J. Shotton, C. Silva, N. Sousa, N. Subbanna, G. Szekely, T. Taylor, O. Thomas, N. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging, 34(10):1993-2024, 2015. 3
- [24] F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision*, *3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 565–571. IEEE Computer Society, 2016. 2
- [25] W. Nawaz, S. Ahmed, A. Tahir, and H. A. Khan. Classification of breast cancer histology images using alexnet. In *International Conference Image Analysis and Recognition*, pages 869–876. Springer, 2018. 6
- [26] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, pages 1520–1528. IEEE Computer Society, 2015. 2
- [27] O. Orzan, A. Sandru, and C. Jecan. Controversies in the diagnosis and treatment of early cutaneous melanoma. *Journal* of medicine and life, 8(2):132, 2015. 1
- [28] J. J. Randolph, A. Thanks, R. Bednarik, and N. Myller. Author note: Free-marginal multirater kappa (multirater kfree): An alternative to fleiss fixed- marginal multirater kappa, 2005. 6
- [29] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. *IEEE Transactions on Biomedical Engineering*, 62(10):2421–2433, 2015. 2
- [30] M. Thoma. A survey of semantic segmentation. arXiv preprint arXiv:1602.06541, 2016. 5
- [31] H. Xu, C. Lu, R. Berendt, N. Jha, and M. K. Mandal. Automated analysis and classification of melanocytic tumor on skin whole slide images. *Comp. Med. Imag. and Graph.*, 66:124–134, 2018. 2
- [32] H. Xu and M. Mandal. Efficient segmentation of skin epidermis in whole slide histopathological images. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 3855–3858, 2015. 6

- [33] X. Xu, Q. Lu, L. Yang, X. S. Hu, D. Z. Chen, Y. Hu, and Y. Shi. Quantization of fully convolutional networks for accurate biomedical image segmentation. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 8300–8308. IEEE Computer Society, 2018. 2
- [34] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5