Automated segmentation of epithelial tissue in prostatectomy slides using deep learning

Wouter Bulten[1,2], Christina A. Hulsbergen - van de Kaa[2], Jeroen van der Laak[1,2], and Geert J.S. Litjens[1,2]

[1]Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, The Netherlands
[2]Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands

ABSTRACT

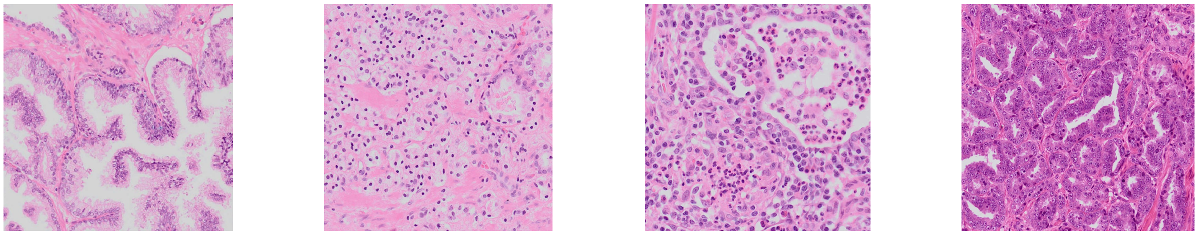
Prostate cancer is generally graded by pathologists based on hematoxylin and eosin (H&E) stained slides. Because of the large size of the tumor areas in radical prostatectomies (RP), this task can be tedious and error prone with known high interobserver variability. Recent advancements in deep learning have enabled development of automated systems that may assist pathologists in prostate diagnostics. As prostate cancer originates from glandular tissue, an important prerequisite for development of such algorithms is the possibility to automatically differentiate between glandular tissue and other tissues. In this paper, we propose a method for automatically segmenting epithelial tissue in digitally scanned prostatectomy slides based on deep learning. We collected 30 single-center whole mount tissue sections, with reported Gleason growth patterns ranging from 3 to 5, from 27 patients that underwent RP. Two different network architectures, U-Net and regular fully convolutional networks with varying depths, were trained using a set of sparsely annotated slides. We evaluated the trained networks on exhaustively annotated regions from a separate test set. The test set contained both healthy and cancerous epithelium with different Gleason growth patterns. The results show the effectiveness of our approach given a pixel-based AUC score of 0.97. Our method contains no prior assumptions on glandular morphology, does not directly rely on the presence of lumina, and all features are learned by the network itself. The generated segmentation can be used to highlight regions of interest for pathologists and to improve cancer annotations to further enhance an automatic cancer grading system.

Keywords: histopathology, whole-slide imaging, deep learning, segmentation, prostate cancer

1. INTRODUCTION

Prostate cancer (PCa) is the most common cancer in men in developed countries.[1] PCa develops from genetically damaged glandular epithelium, resulting in altered cellular proliferation patterns. In the case of high-grade tumors, the glandular structure is eventually lost.[2] Treatment planning is generally based on histological grading of prostate biopsies (preoperative) or full radical prostatectomy (RP) slides (postoperative). The Gleason score is the most important marker for patient prognosis and is determined by pathologists on H&E stained specimens. As PCa originates from epithelial cells, glandular structures within prostate specimens are regions of interest for finding malignant tissue. For a pathologist, assessing all epithelial regions can be a time-consuming task, especially when considering the gigapixel-sized RP slides, the poorly differentiated structure of high-grade PCa and the heterogeneity in prostate cancer growth patterns. An automated method to highlight these regions can help speed up this task.

Moreover, automatically differentiating between glandular tissue and other tissues is an important prerequisite for the development of automated methods for detecting PCa. Typically, deep learning methods that try to detect cancer from scanned tissue specimens use a set of annotated cancer regions as the reference standard for training. As these algorithms learn from their training data, the quality of the annotations directly influences the quality of the output. Outlining all individual tumor cells within PCa is practically infeasible due to the mixture of glandular, stromal and inflammatory components (Figure 1). Therefore, tumor annotations made by pathologists often contain large amounts of non-relevant tissue, which adds noise to the reference standard and, subsequently, limits the potential of these deep learning methods. By automatically removing all non-relevant tissue, these coarse tumor annotations can be refined and the ability of these networks to detect cancer could be improved.



(a) (b) (c) (d)

Figure 1: Different types of glands: normal glandular structure (1a); poorly differentiated, high-grade Gleason 5 PCa (1b); non-tumor epithelium surrounded by inflammation (1c); Gleason 3 PCa showing color variation between slides (1d).

Given a set of labels, in this case epithelial and non-epithelial tissue, we trained a system to automatically divide digitized tissue into relevant and non-relevant tissue on a pixel-by-pixel basis. Existing research shows promise in this task for PCa: Gertych et al.[3] use texture and intensity based features in combination with an SVM; Naik et al.[4] apply a Bayesian classifier, relying on the detection of lumina; Singh et al.[5] propose a multi-step solution including nuclei segmentation to create a final segmentation map of glandular regions. While these approaches achieve a good performance, they are often only trained or tested on low grade PCa or rely on the morphological structure of the glands. The wide-ranging glandular structures in PCa, such as the cribriform growth pattern of Gleason 4, limit the applicability of these methods. By applying deep learning we can try to overcome this problem by not imposing any predefined features on the model. Instead, we rely on the network itself to learn the relevant features directly from data. Previous research shows that this task is possible on other tissue types, e.g. in breast and colorectal cancer.[6]

In this research we propose an automated method for segmentation of epithelial tissue within prostatectomy slides using deep convolutional neural networks (CNN). Using CNNs we remove the need for predefined features. We compare two different architectures: regular fully-convolutional networks[7] and U-Net.[8] Our automated segmentation is not only useful as a tool for pathologists, we also envision this segmentation as the first part of a fully automated prostate cancer detection and grading pipeline.

2. METHODOLOGY

2.1 Multi-resolution histopathology images

We collected 30 single-center whole mount tissue sections from 27 patients that underwent RP treatment. The reported Gleason growth patterns in these sections ranged from 3 to 5. The specimens were prepared with a H&E stain, subsequently digitized and randomly split into three sets: 15 slides for training, 5 for validation and 10 for testing. Within the training and validation slides epithelial structures and stroma regions where delineated by a non-expert under supervision of an experienced pathologist, specialized in uropathology. Lumina were removed from the annotations using a color-based background filter and further refined by hand. An initial version of our system was applied to the training slides and a second annotation round was used to annotate regions that were initially misclassified by the network.

The total number of epithelium annotations per slides are low; in total 15-30 different regions were annotated per slide which resulted in an average annotation coverage of approximately 0.03% per slide. We trained our networks patch-by-patch and these patches were sampled from the slides at 10x magnification (pixel resolution 0.64 μm).

2.2 Network architecture

Two types of networks were trained: a regular fully convolutional network (FCN)[7] and U-Net.[8] FCNs are the de-facto standard for deep learning on this type of data, whereas U-Nets have been specifically designed for segmentation. We compare these methods to determine whether U-Net offers tangible benefit over FCNs for

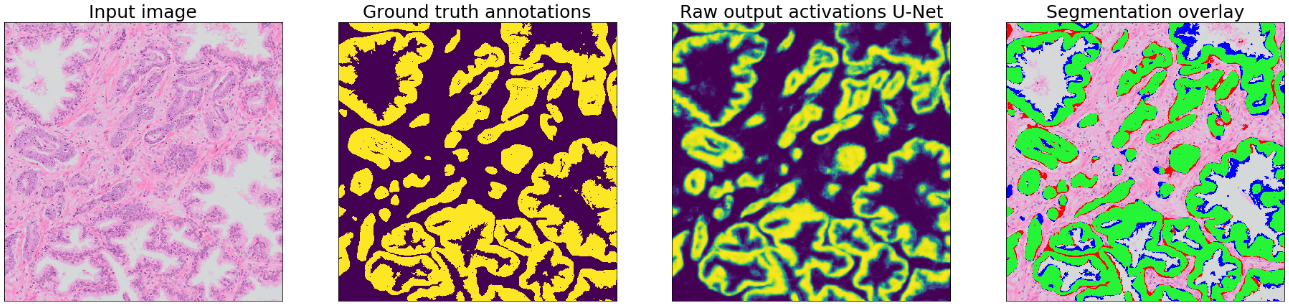


Figure 2: Overview of the evaluation method using the output of the U-Net depth 4 as an example. For each input image from the test set, ground truth annotations were made manually by hand. The network is applied to the input image and the raw output of the network is thresholded to generate a binary segmentation map. The segmentation map is then compared with the ground truth. The segmentation overlay shows the performance of the network: green marked pixels show true positive, blue false negative and red false positive.

this specific problem. Besides the network architecture, we also vary the complexity of both models by testing different settings for the network depth. Both networks were implemented using Theano[9] and Lasagne.[10]

The FCN has an input size of $(128 \times 128 \times 3)$ (width, height, channels) and an output of (1×2) (one output for each class). During training the FCN classifies the central pixel of each patch. Each FCN network consists of a number of contraction blocks defined by the *depth* parameter, and two classification layers. Each contraction block consists of two convolution layers and a max pooling layer. For the last contraction block the pool layer is omitted. The classification layers consist of two (1×1) convolutional layers with respectively 2048 and 1024 filters that mimic the behavior of a fully connected layer. The depth of the network was varied from 2 to 4. Given d the depth of the current layer and D the total depth of the network, the network structure is as follows:

$$\begin{bmatrix} \text{Conv-}3\times 3 & 2^{d+4} \\ \text{Conv-}3\times 3 & 2^{d+4} \\ \text{MaxPool-}2 & \end{bmatrix} \times (D-1) \begin{bmatrix} \text{Conv-}3\times 3 & 2^{d+4} \\ \text{Conv-}3\times 3 & 2^{d+4} \end{bmatrix} \begin{bmatrix} \text{AvgPool} \end{bmatrix} \begin{bmatrix} \text{Conv-}8\times 8 & 2048 \\ \text{Conv-}1\times 1 & 1024 \end{bmatrix} \begin{bmatrix} \text{Dropout} & \\ \text{Conv-}1\times 1 & 2 \end{bmatrix}$$

An average pool layer is used to match the output of the contraction layers to the input of the classification layers (and omitted in the depth 4 network). Batch normalization was applied to all convolution layers.

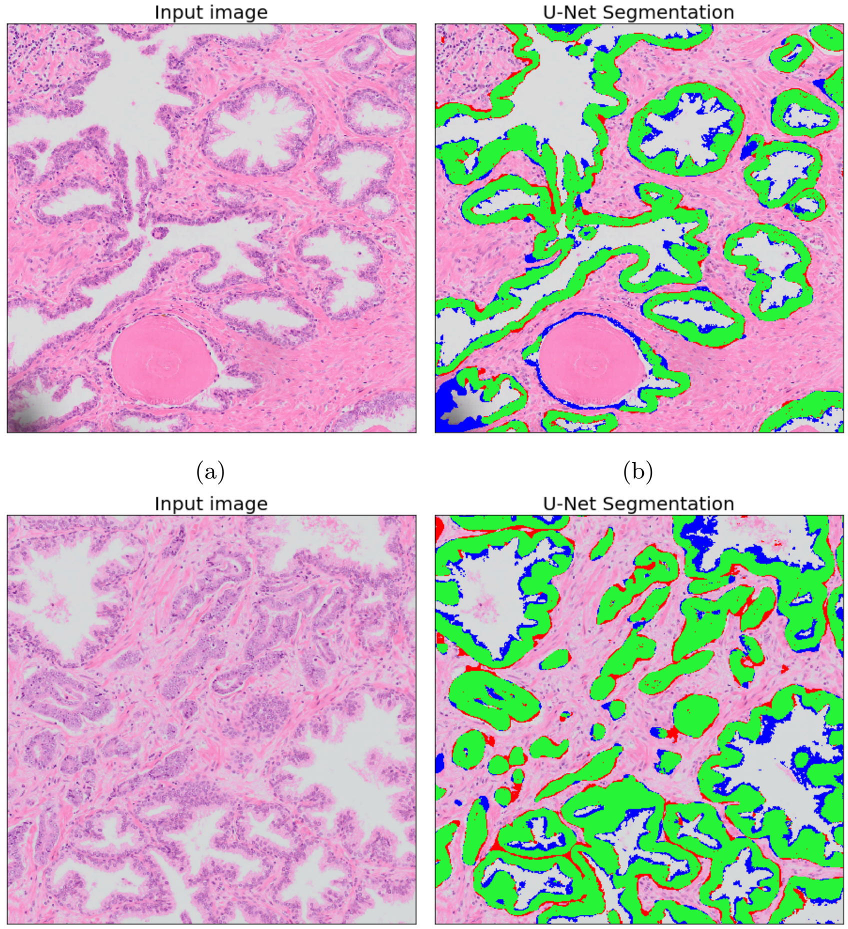
For our U-Net implementation we followed the architecture by Ronneberger et al.[8] The U-Net architecture makes use of two paths, a contracting and expansive, and outputs a segmentation map (in contrast to the FCN). Our U-Net networks uses input patches of $(512 \times 512 \times 3)$. Three different depths were tested. The complexity of the contracting path, in terms of number of parameters, was equal to the complexity of the FCN (see also Table 1).

2.3 Evaluation method

For each test slide, the region of the primary tumor was delineated by a pathologist if the slide contained PCa. From each test slide, we randomly selected two regions of (1250×1250) pixels each, one from a cancer and a second from a non-cancer area, at 10x magnification. If no cancer was present in the slide, two non-cancer regions were selected. A new random region was selected if there were major scanning artifacts present, or when there was no epithelium or cancer tissue in the region (e.g when a primarily stromal area within a cancerous region was selected).

In total we sampled 8 cancer and 12 non-cancer regions. After selection, all epithelial tissue within these regions was annotated exhaustively. Lumina were removed with a background filter and further refined by hand. See Figure 2 for an overview of the evaluation method.

After training, the optimal classification threshold was determined for each network individually based on the validation set. The trained networks were then applied to the individual regions of the test set.



(a) (b)

(c) (d)

Figure 3: Best performing network, U-Net depth 4, applied to two regions from the test set. Green marked pixels show true positive, blue false negative and red false positive. Most errors are present at the border of the epithelium which are in some cases caused by errors in the reference standard. Figure 3a shows an example of an artifact in the bottom left corner. The network does not rely on the presence of lumen as can be seen in Figure 3d.

3. RESULTS

The thresholded output of the networks was compared with the ground truth. For each network the pixel-based performance was computed using accuracy, F1 and Jaccard score (Table 1). Examples of the final segmentation output can be seen in Figure 3.

The networks with the highest depth setting achieved the highest accuracy. The best FCN and U-Net achieve F1 scores of 0.83 and 0.82 with a total AUC of 0.96 (Fig. 5a) and 0.97 (Fig. 5b) respectively. While the performance of both network architectures is comparable, the U-Net is more efficient in terms of parameter complexity and performs slightly better on the cancer regions (AUC of 0.95 versus 0.92, see Figure 6 for an example).

4. DISCUSSION

We tested two different deep learning architectures, a regular fully convolutional network and U-Net, to segment epithelium from H&E stained prostatectomy slides. Our approach shows that it is possible to detect epithelium/glandular structures without relying on any a priori defined features or domain knowledge embedded in the

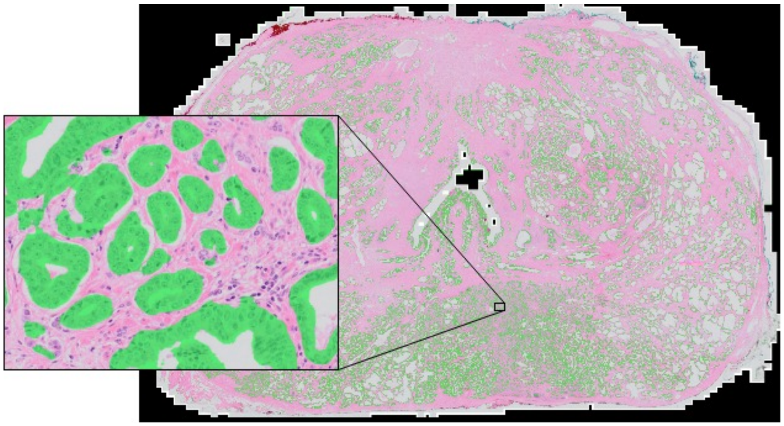


Figure 4: Example of applied network on whole-slide level. Although the networks are trained patch-by-patch, the goal is to apply the networks on a whole-slide level. The thresholded network output is shown in green.

Table 1: Network overview, complexity and pixel-based performance on the test set. The number of contraction parameters is based on the layers from the downward path for the U-Net and the and the contraction blocks of the FCN. The total number of parameters is based on all the layers of the network. F1, Jaccard and accuracy scores are based on the whole test set. The cancer and non-cancer AUC were solely calculated on the test regions that contained cancer or no cancer respectively. Highest scores are marked in bold.

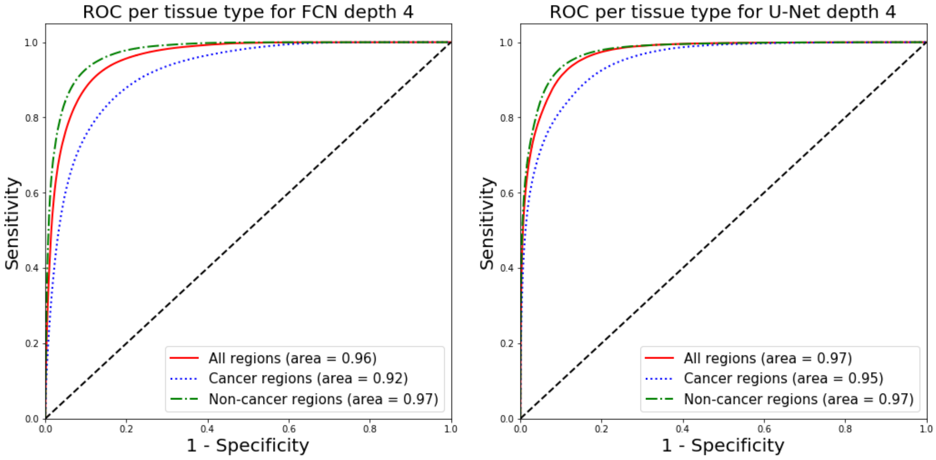
#	Network	Depth	Contraction parameters	Total parameters	F1	Jaccard	Accuracy	AUC total	AUC cancer	AUC non-cancer
1	U-Net	2	16,944	26,130	0.79	0.66	0.87	0.95	0.93	0.94
2	U-Net	3	72,752	118,162	0.82	0.70	**0.90**	0.96	**0.95**	0.96
3	U-Net	4	294,960	484,498	0.82	0.70	**0.90**	**0.97**	**0.95**	**0.97**
4	FCN	2	16,944	6,322,738	0.79	0.66	0.87	0.94	0.90	0.96
5	FCN	3	72,752	10,572,850	0.81	0.69	0.88	0.95	0.90	**0.97**
6	FCN	4	294,960	19,183,666	**0.83**	**0.71**	0.89	0.96	0.92	**0.97**

model. In terms of layer depth, the deepest version of both architectures perform best on our test set. Given the metrics there is no clear winner in terms of segmentation performance as both models achieve high scores. Though, taking parameter complexity in to account, the U-Net outperforms the FCN as it achieves a comparable accuracy using only a fraction of the parameters.

It is hard to compare our results to existing methods due to the lack of a shared test set and varying complexity in terms of Gleason growth patterns. Moreover, where we focus specifically on segmenting epithelial tissue others include the lumen in the segmentation. Apart from these limitations in the comparison, our method seems to outperform existing gland segmentation approaches; e.g. Singh et al.[5] (F-score of 0.74) and Gertych et al.[3] (Jaccard score of 0.595).

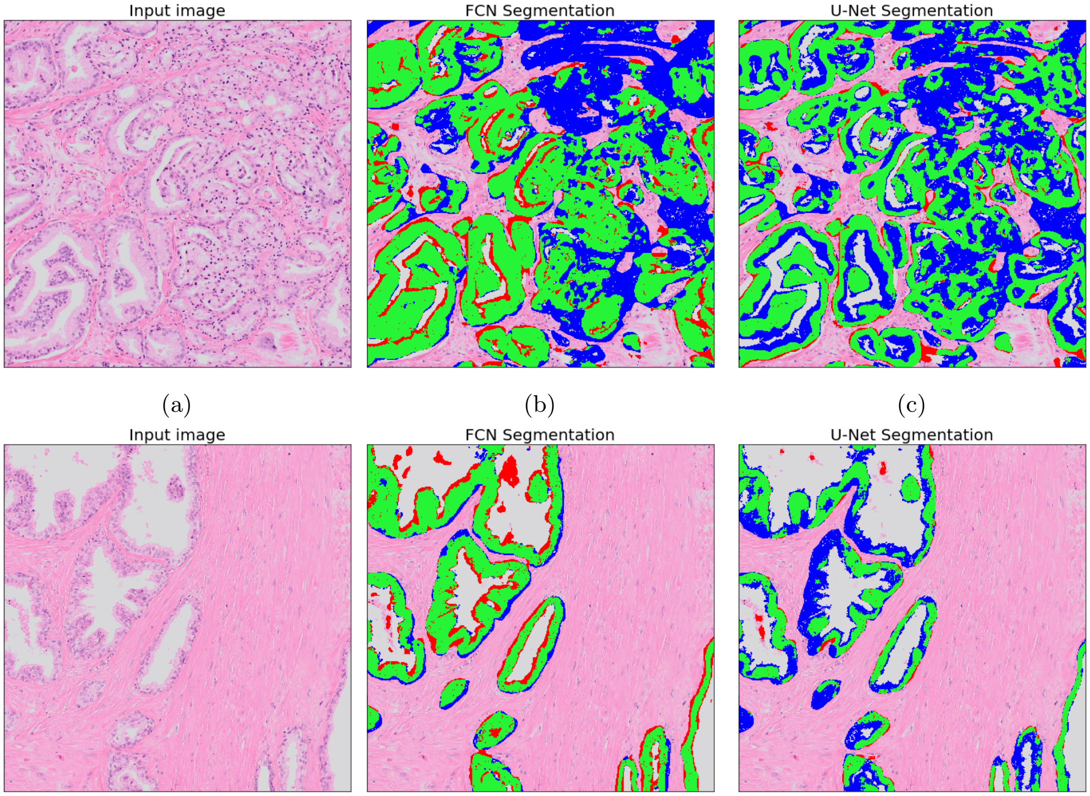
Room for improvement lays with segmenting glands as a whole. Our models rarely miss complete cancer regions, though with some high-grade PCa only parts of the glands are detected (Figure 6c). We suspect that most of the errors are, first of all, caused by a lack of training examples and not due to a limitation of the models. In our dataset, the occurrence of high grade PCa (growth patterns Gleason 4 and Gleason 5) was lower than the more common low grade PCa (Gleason 3) or healthy epithelium. Moreover, these high grade tumors are often poorly differentiated which makes manual annotating these glands difficult. Even for a trained expert it can be hard to precisely outline epithelial cells in H&E, especially in regions with inflammation or when a gland is deformed. The border of the epithelium, especially within the lumen, is also not always clearly defined.

We want to address the limitations of our dataset by collecting more slides that contain high grade PCa. We expect that our models will perform better if we include more specimens of diverse and high grade tumors.



(a) (b)

Figure 5: ROC curves for best performing FCN (5a) and U-Net (5b), each network has a depth of 4. Both networks perform less on the cancer regions in comparison to the non-cancer regions. The U-Net performs slightly better on the cancer regions but this difference is very small.



(a) (b) (c)

(d) (e) (f)

Figure 6: Example of failure cases where one of the networks outperforms the other. The top row contains PCa (Gleason growth pattern 4) whereas the bottom row only contains benign epithelium. In the cancer example, the U-Net outperforms the FCN by segmenting more of the Gleason 4 PCa whereas the FCN misses this almost completely (right top, 6a). In the non-cancer example,tThe U-Net (6f) under-segments while the FCN (6e) finds more of the epithelium at the cost of over-segmenting portions of the lumina.

Including more high grade PCa in our dataset has the downside that making annotations will be increasingly difficult. Our current dataset suffers from this problem and any larger set will face the same problem. To create a more precise ground truth we plan to use immunohistochemistry to assist in making annotations. By using specific stains that highlight epithelial cells, we hope to generate precise and less erroneous annotations of epithelial tissue.

REFERENCES

[1] Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-tieulent, J., and Jemal, A., "Global Cancer Statistics, 2012," *CA: a cancer journal of clinicians.* **65**(2), 87–108 (2015).

[2] Fine, S. W., Amin, M. B., Berney, D. M., Bjartell, A., Egevad, L., Epstein, J. I., Humphrey, P. A., Magi-Galluzzi, C., Montironi, R., and Stief, C., "A contemporary update on pathology reporting for prostate cancer: Biopsy and radical prostatectomy specimens," *European Urology* **62**(1), 20–39 (2012).

[3] Gertych, A., Ing, N., Ma, Z., Fuchs, T. J., Salman, S., Mohanty, S., Bhele, S., Velásquez-Vacca, A., Amin, M. B., and Knudsen, B. S., "Machine learning approaches to analyze histological images of tissues from radical prostatectomies," *Computerized Medical Imaging and Graphics* **46**, 197–208 (2015).

[4] Naik, S., Doyle, S., Feldman, M., Tomaszewski, J., and Madabhushi, A., "Gland Segmentation and Computerized Gleason Grading of Prostate Histology by Integrating Low- , High-level and Domain Specific Information.," in [*Proceedings of 2nd Workshop on Microsopic Image Analysis with Applications in Biology*], 1–8 (2007).

[5] Singh, M., Kalaw, E. M., Giron, D. M., Chong, K.-T., Tan, C. L., and Lee, H. K., "Gland segmentation in prostate histopathological images.," *Journal of medical imaging* **4**(2), 027501 (2017).

[6] Xu, J., Luo, X., Wang, G., Gilmore, H., and Madabhushi, A., "A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images," *Neurocomputing* **191**, 214–223 (2016).

[7] Shelhamer, E., Long, J., and Darrell, T., "Fully Convolutional Networks for Semantic Segmentation," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 3431–3440 (2015).

[8] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*], Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., eds., 234–241, Springer International Publishing, Cham (2015).

[9] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints* **abs/1605.02688** (May 2016).

[10] Dieleman, S., Schlter, J., Raffel, C., Olson, E., Snderby, S. K., Nouri, D., et al., "Lasagne: First release.," (Aug. 2015).