CS231n Final Report

Pierre Louis Maurice Valentin Cedoz
Stanford
Center for Biomedical Informatics Research
plcedoz@stanford.edu

Quinlan Jung
Stanford
Computer Science
quinlanj@stanford.edu

Olivier Gevaert
Stanford
Center for Biomedical Informatics Research
ogevaert@stanford.edu

Abstract

DNA methylation is an important mechanism regulating gene transcription, and its role in carcinogenesis has been extensively studied. Hyper and hypomethylation of genes is a major mechanism of gene expression deregulation in a wide range of diseases. At the same time, high-throughput DNA methylation assays have been developed generating vast amounts of genome wide DNA methylation measurements. However, these assays remain extremely expensive and they are not performed systematically. On the contrary, pathology images are part of the common procedure in cancer treatment and they are relatively cheap. The goal of this project was to study the correlations between pathology images and DNA methylation. We have shown that it was possible to predict the methylation profile of a patient from the pathology images using deep learning. More precisely, we used convolutional neural networks on pathology images to predict the methylation state of the patient (binary value). This project idea originated from Pr Gevaert's group at Stanford who developed a software called MethylMix to identify disease specific hyper and hypomethylated genes.

1. Introduction

DNA methylation is one of the most studied epigenetic aberrations underlying oncogenesis. Besides genetic mutations, hyper and hypo- methylation of genes is an alternative mechanism that is capable of altering the normal state and driving a wide range of diseases. Several computational tools have been developed incorporating state-of-the-art statistical techniques for the analysis of DNA methylation data. However, obtaining a patients methylation profile is extremely expensive. Plus, whenever a patient is suspected to have cancer, multiple histology sections are taken systematically and the cost is relatively low. The idea behind this project was to use state-of-the-art machine learning techniques to predict the methylation state of a cell (hyper or hypo methylated) from histology sections of a patients tissue. This is a very significant problem since our algorithm could avoid the burden of DNA-methylation profiling and extract all relevant information from the pathology images.

The project was divided into three main parts:

1. Build a preprocessing pipeline to load the pathology images and prepare them to be inputted into a deep learning model.
2. Implement a deep convolutional neural network using transfer learning.
3. Design an evaluation metric for our models.

2. Related Work

In a recent paper [3], *Rubin et al* tackled the problem of detecting the presence of brain cancer by training a CNN on the same pathology images dataset with label the type and stage of cancer. One of the problems in classifying histology sections is dealing with a high resolution image with sparse cell clusters. Examining the textures of the cell clusters is the most important factor in determining whether there are any issues.

An other group [1] worked on multiple instance learning with pathology images. Since pathology images contain multiple cell instances (as opposed to traditional classification problems where there is only one instance), they modified the common CNN framework by replacing pooling layers by the aggregation function used in multiple instance learning.

Layer	Type	Num Kernels	Kernel size	Stride	Activation
0	Input	3	32×32	-	-
1	Convolution	32	5×5	1	-
2	Max pool	-	3×3	2	ReLu
3	Convolution	32	5×5	1	ReLu
4	Mean pool	-	3×3	2	
5	Convolution	64	5×5	1	ReLu
6	Mean pool	-	3×3	2	
7	Fully connected	64	-	-	Dropout+ ReLu
8	Fully connected	2	-	-	Dropout+ ReLu
9	SoftMax	-	-	-	

Figure 1. Janowczyk et al's neural network configuration, using the AlexNet configurations except layers 7 and 8 of the dropout network have an additional dropout combined with the ReLu.

Variable	Setting
Batch size	128
Initial learning rate	0.001
Learning rate schedule	Adagrad
Rotations	0, 90
Number of iterations	600,000
Weight decay	0.004
Random minor	Enabled
Transformations	Mean-centered

Figure 2. Janowczyk et al's deep learning hyperparameter settings held constant for all experiments.

Janowczyk et al [19] classify digital pathology images by using transfer learning with a modified version of Alexnet as shown in Figure 1. They were able to achieve extremely good results with a neural network that had very similar parameters to Alexnet, showing how parameter tweaking and tuning is not strictly necessary to yield good quality results, as shown in Figure 2. Transfer learning is currently the state-of-the-art approach for classifying tissue images, with accuracy scores of $>85\%$ on small data sets ($n < 200$) of high resolution images [20] [22] [24] [23], and $>93\%$ on larger datasets ($n > 400$). [21] [25]

Nuclei segmentation is an effective technique in tissue classification tasks that improves classification accuracy. [22] There is evidence that the configuration of nuclei is correlated with outcome, and nuclear morphology is a key component in most cancer grading schemes. [27] By creating a bitmap image denoting nuclei and non-nuclei regions, classification accuracy improves. To do this, patches are selected from the positive class, and a threshold is used on the color-deconvolved image to determine examples of non-nuclei regions. [26] One weakness of this method is that the initial manual classification of nuclei is laborious and time consuming. Currently, the state of the art technique is to use another convolutional neural network to classify nuclei regions [30] [29], which has yielded 96.7% true positive and 94.2% true negative scores. [28]

3. Method

3.1. Data preprocessing pipeline

When using deep learning on whole slide images, the particularity is that the images are very wide (typically 100,000 pixels wide) so it is impossible to feed them directly into the CNN because of memory constraints. They need to be divided into smaller patches (up to 10,000 patches for a single image). This was the strategy adopted in a recent paper from Google where they present their work in the Camelyon16 challenge that involved detecting small tumors in gigapixel pathology slides [2]. We will adopt their implementation method of detecting where the cells are in each section and extracting smaller images of the cell textures. These smaller images will be the input to our CNN.

Each image from the TCGA dataset is around 26000x26000 pixels, and since it is impractical to put such a large image through a CNN, divide each image into 896x896 tiles. Each image contains a tissue sample on a slide background, with the tissue being the object of interest. In order to give the CNN the most relevant input, we filter out the tiles that do not contain at least 90% tissue.

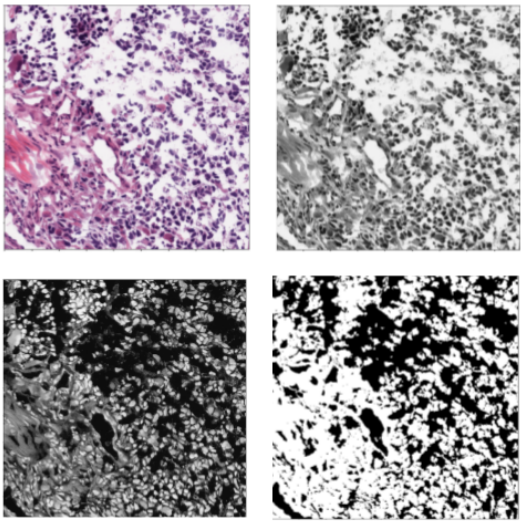

Figure 3. Top Left: An unprocessed patch of tissue, Top Right: Greyscale applied, Bottom Left: 8 bit complement applied, Bottom Right: Hysteresis thresholding of 0.5 applied. If a patch has less than 0.5 white pixels, we discard it.

We adopt the Rubin et al's method of tissue filtering to distinguish tissue from the slide background. We apply the

following procedure to each tile:

1. Apply a greyscale filter. We use a nonlinear luma component (Y') that is calculated directly from gamma-compressed primary intensities as a weighted sum, which can be calculated quickly without the gamma expansion and compression used in colorimetric grayscale calculations. The following equation is used to compute greyscale: $Y' == 0.299R + 0.587G + 0.114B$.

2. Take the 8-bit image complement as follows: $I' = 255 - I$, where I is the 2D greyscale image matrix.

3. Perform hysteresis thresholding with an experimentally-chosen high threshold of 100 and a low threshold of 50.

3.2. Preprocessing runtime

Processing whole slide images (WSI) is a computationally expensive and time consuming task. In order to preprocess 117 slide images, Ruiz et al [18] focused their work on using a GPU to reduce the execution time. We preprocess 275 slide images in 3 hours using multithreading and program optimization.

3.2.1 Multithreading

We use a n1-highmem-8 (8 vCPUs, 52 GB memory) Google Compute machine to process each image sequentially. The WSI's are divided amongst 8 threads, where the number of threads in the pool was experimentally found to have the lowest runtime. Multithreading yields a 4.5x speedup from our naive solution.

3.2.2 Program Optimization

We optimize our program to run as fast as possible, avoiding expensive operations like numpy append, which creates new copies of objects. Also, we also modify Rubin et al's filtering methods to exclude computationally expensive steps like autocontrasting, opting to change the parameters of our other filtering steps instead to compensate for the missing step. Program optimization yields a 3x speedup from our naive solution.

3.3. Deep learning approach

3.3.1 Convolutional Neural Networks

Convolutional Neural Networks are a type of discriminative deep architecture in which layers consisting of a convolutional layer and a pooling layer are often stacked up with one on top of another to form a deep model [6]. The convolutional layers share many weights, and the pooling layers subsample the output of the convolutional layers and reduce the data rate from the layer below. The weight sharing in the convolutional layer, together with appropriately chosen pooling schemes, endows the CNN with some invariance properties (e.g., translation invariance). A series of fully connected layers is stacked on top of the last convolutional layer and finally a classifier (SVM or softmax) is applied to classify the image. In our case, there are two target labels: hyper-methylated or not (**Figure 1**)

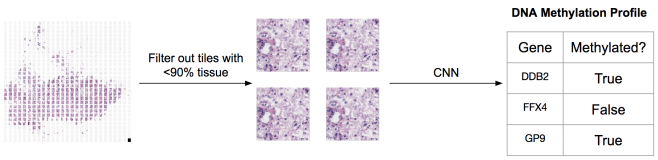

Figure 4. Convolutional Neural Network trained to recognize the methylation value of a particular gene.

CNNs have been found highly effective and been commonly used in computer vision and image recognition [7]. They have the advantage of automatically learning the appropriate features, as opposed to traditional machine learning approaches that use hand-crafted features. CNN architectures extract local features at a high resolution and successively combine these into more complex features at lower resolutions.

3.3.2 Transfer Learning

Convolutional Neural Networks are very computationally expensive to train in practice and they require a large dataset to avoid overfitting. Therefore, we applied transfer learning concepts by using a pretrained Convolutional Neural Networks on the ImageNet dataset. More precisely, we started from a network pre-trained on ImageNet, removing the last fully-connected layer, then treating the rest of the ConvNet as a fixed feature extractor for the pathology images dataset. We experimented with the InceptionV3 [15] and ResNet [14] architectures that computed a vector representation for every image immediately before the classifier. Then, we trained a new classifier on the pathology images dataset (**Figure 2**). We experimented with different classifier architectures: number of fully connected layers and number of neurons. Since our dataset was not very large, we also incorporated dropout layers, which is a regularization technique developed by Srivastava et al. [10]. We didn't fine-tune the weights of the pre-trained network because our dataset was too small and prone to over-fitting.

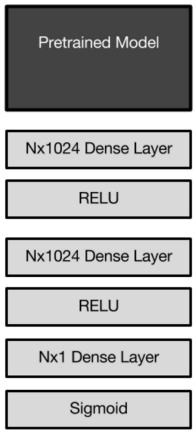

Figure 5. Transfer Learning approach: A new classifier was trained on top of the last layer of the pre-trained model

3.4. Training and Evaluation

3.4.1 Deep Learning Framework

Our deep learning framework was Keras, which is a high-level neural networks API, written in Python and capable of running on top of either TensorFlow or Theano. We built our convolutional neural network using the pretrained model InceptionV3 in Keras Model Zoo and stacking the classifier discussed in the previous section on top of it. Then we used an optimization algorithm to train our network called Adam [13] that is a variation of the traditional gradient descent algorithm.

3.4.2 Evaluation Metrics

Our model contained many hyperparameters: the learning rate, the number and type of layers, the regularization parameters and the optimizer. A 30 % hold-out validation technique was used to train our model and tune these hyperparameters. Then we evaluated our best model on an independent set of pathology images. We tackled a multi-task, multi-class classification problem. Each task corresponded to a gene and each class corresponds to a methylation state (hyper, hypo or null). To start with, we focused on one gene and used the cross-entropy loss function to train and evaluate our model:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \qquad (1)$$

where N is the number of images in the test set, M is the number of categories, log is the natural logarithm, yij is 1 if observation i belongs to class j and 0 otherwise, and pij is the predicted probability that observation i belongs to class j.

3.4.3 Training and Testing accuracies

Since the loss score was not very interpretable, we computed the accuracy of our models as an indicator of their performance. The evaluation procedure was different during training and testing. During training, we assigned to every patches in every images the label of the global images and we trained a network to classify the patches. Therefore, we computed loss and accuracies at the patch level. However, during testing, we split every test images into patches, classifying every patches in the image and using a majority voting procedure to classify the whole image. The accuracy metrics were then computed at the global image level.

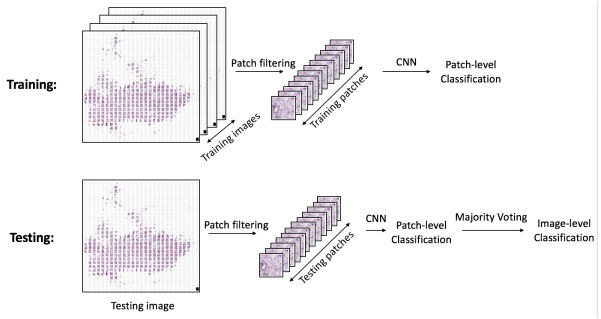

Figure 6. Evalution procedures. **Top:** patch-level classification at training time. **Bottom:** image-level classification at testing time.

4. Experiments

4.1. Dataset

We downloaded our datasets from the open source The Cancer Genome Atlas (TCGA) data portal, which is a repository of molecular, clinical and biological data for all cancer sites. We decided to focus on Glioblastoma Multiforme (GBM) and Lower Grade Glioma (LGG) cancer types because these cancers are the focus of Pr Gevaert's lab and they are very aggressive. We downloaded the pathology images for each patients in this cohort (1121 patients) and labeled each picture with the patients methylation profile for every gene. This methylation profile was computed using a Software called MethylMix developed in Pr Gevaerts Lab.

4.2. Preprocessing results

The original tissue filtering algorithm included contrast stretching the greyscale image. This uses upper and lower pixel value limits over which the image is to be normalized. However, we found that contrast stretching produced many false positives since there were some background slide tiles that became significantly darker and was mistaken as tissue when performing hysteresis thresholding. To compensate, we discovered that setting the hysteresis threshold to 50% yielded similar results that Rubin et al obtained.

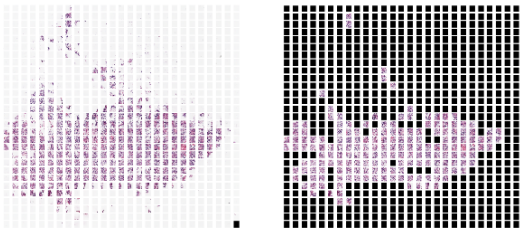

Figure 7. Whole Slide Image after preprocessing. Left: A typical cell section, Right: A filtered image. Black tiles do not contain enough tissue, so they are not included as inputs in our neural network.

These preprocessing steps were very slow (10 min per image), so we parallelized the code to run on multiple CPUs to process the entire dataset.

4.3. Deep Learning results

Given the fact that the training procedure is different than the testing procedure, we developed different evaluation metrics for these two phases.

4.3.1 Results on training

During training, we assigned each patch the label of the global image, and we trained a network to classify the patches. Therefore, we computed loss and accuracies at the patch level. This is a classical configuration for deep learning pipelines and we plotted the learning curves for our best model (**Figure 5 and 6**): ResNet architecture with learning rate = 1e-4, Beta1 = 0.9, Beta2 = 0.999, epsilon = 1e-08, decay = 1e-6

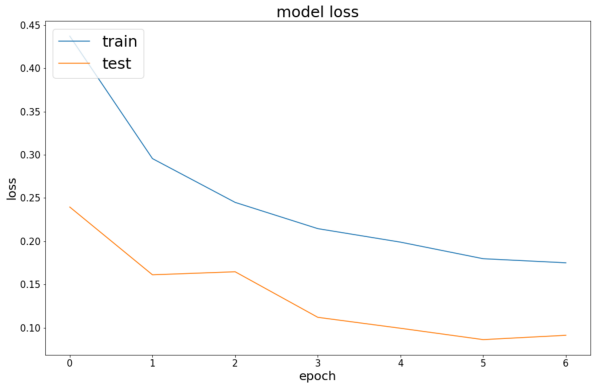

Figure 8. Learning curve for the best model: log-loss over the number of epochs

We observe that the test accuracy is higher than the training accuracy and that the testing loss is lower than the training loss because of the random effects in the dropout layer during training that become deterministic at test time. Overall, our model performs very well and reaches a classification accuracy of 95.8% at the patch level. This is very impressive because we only trained on 28 images that accounted for about 3500 patches. However, we need to assess the generalization power of our model using an independent set of images.

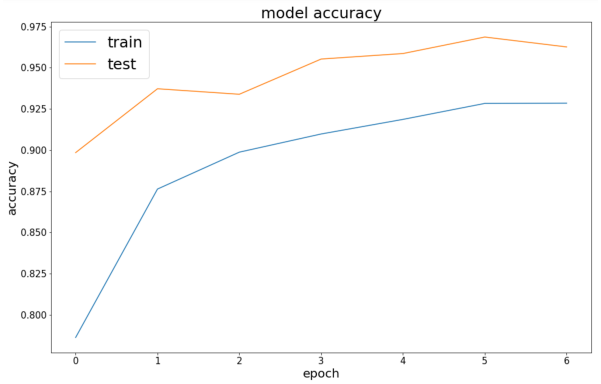

Figure 9. Learning curve for the best model: accuracy over the number of epochs

4.3.2 Results on testing

During testing, we computed the classification accuracy of our models on the test set consisting of 12 independent images. We split every test image into patches, classified every patches in the image and used a majority voting procedure to classify the whole image. The accuracy metrics were then computed at the global image level.

We obtained an accuracy of 84% with the ResNet architecture and 75% with the Inception architecture. This is a proof of concept of our assumption that the methylation state of a gene can be predicted from the pathology images.

4.3.3 Tile resolution

We also experimented with different tile resolutions. The high resolution is 20x corresponding to the original resolution of the whole slide image, the low resolution is 5x. There is a trade-off between the number of nuclei seen on the image and the magnification of these nuclei. We realized that the model yielded much better performance with the high resolution patches. In the high resolution patches, there are fewer nuclei but they are more defined so this experiment confirms that the nucleus morphometric features are very important in the prediction task [16].

5. Conclusion and Perspectives

For next steps, we will train the algorithm on a larger dataset of available images (n=1000). With a larger dataset, we would also experiment with finetuning larger parts of Resnet/Inception in our current neural network to see if we can achieve higher levels of accuracy.

Both the results in our experiment and the prior literature show that examining a cell's nuclei is crucial to determine the presence of methylation [17]. If we segmented the nuclei from the patches using a series of image complements and experimentally determined hysteresis thresholding values, we could input the segmented patches into the network. However, the state of the art technique in nuclei segmentation is constructing another CNN to classify nuclei regions, so it would also be an engineering challenge to have such a model perform at scale (ie) optimize runtime when classifying over 250 GB of slide images.

Our experiments show that we can predict the methylation values of a single gene with relatively high accuracy. Originally, our goal was to predict a complete methylation profile (ie) given an image, predict the presence/absence of methylation in n genes. Given more time, instead of predicting the methylation value of a single gene, we would instead be outputting a vector, where each element represents the methylation value of a gene. If we are able to predict a methylation profile with similar accuracy from our experiments, we would have created an extremely useful tool. We could help patients avoid the burden of having DNA-methylation profiles being performed in the lab.

References

[1] Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics. 2016;32(12):i52-i59.

[2] arXiv:1703.02442

[3] Ertosun MG, Rubin DL. Automated Grading of Gliomas using Deep Learning in Digital Pathology Images:A modular approach with ensemble of convolutional neural networks. AMIA Annual Symposium Proceedings. 2015;2015:1899-1908.

[4] Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles.

[5] Gurcan et al., 2006 M.N. Gurcan, T. Pan, H. Shimada, J. Saltz Image analysis for neuroblastoma classification: segmentation of cell nuclei Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society1 (2006), pp. 48444847

[6] LeCun Y, Bottou L, Bengio Y, Haffner P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86:22782324.

[7] Deng, Li, and Dong Yu: Deep learning: methods and applications. Foundations and Trends in Signal Processing 7.34 (2014): 197-387.

[8] Nair, Vinod, and Hinton, Geoffrey E.: Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10). 2010.

[9] Ioffe, Sergey, and Christian Szegedy: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015).

[10] Srivastava, Nitish, et al: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15.1 (2014): 1929-1958.

[11] Abadi, Martin, et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016).

[12] Duchi, John, Elad Hazan, and Yoram Singer.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12.Jul (2011): 2121-2159.

[13] Kingma, Diederik, and Jimmy Ba: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[14] He, Kaiming, et al.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

[15] Szegedy, Christian, et al.: Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

[16] Kong J et al.: Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. PLoS One. (2013)

[17] Kong J et al.: In Silico Analysis of nuclei in glioblastoma using large-scale microscopy images improves prediction of treatment response Conf Proc IEEE Eng Med Biol Soc. (2011)

[18] Ruiz et al., 2007 A. Ruiz, O. Sertel, M. Ujaldon, U. Catalyurek, J. Saltz, M. Gurcan Pathological image analysis using the GPU: stroma classification for neuroblastoma Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007. (2007), pp. 7888

[19] Janowczyk, Andrew, and Anant Madabhushi. "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases." Journal of pathology informatics 7 (2016).

[20] Anthimopoulos, Marios, et al. "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network." IEEE transactions on medical imaging 35.5 (2016): 1207-1216.

[21] Aichingera, Wolfgang, et al. "Automated cancer stem cell recognition in H&E stained tissue using convolutional neural networks and color deconvolution." SPIE Medical Imaging. International Society for Optics and Photonics, 2017.

[22] Janowczyk, Andrew, et al. "A resolution adaptive deep hierarchical (radhical) learning scheme applied to nuclear segmentation of digital pathology images." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2016): 1-7.

[23] Huynh, Benjamin Q., Hui Li, and Maryellen L. Giger. "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks." Journal of Medical Imaging 3.3 (2016): 034501-034501.

[24] JimenezdelToroab, Oscar, et al. "Convolutional neural networks for an automatic classification of prostate tissue slides with highgrade Gleason score." Proc. of SPIE Vol. Vol. 10140. 2017.

[25] Chen, Richard, Yating Jing, and Hunter Jackson. "Identifying Metastases in Sentinel Lymph Nodes with Deep Convolutional Neural Networks." arXiv preprint arXiv:1608.01658 (2016).

[26] RuifrokAC, JohnstonDA. Quantification of histochemical staining by color deconvolution.Anal Quant Cytol Histol 2001;23:2919

[27] Basavanhally A, Feldman M, Shih N, Mies C, Tomaszewski J, Ganesan S, et al. Multifieldofview strategy for imagebased outcome prediction of multiparametric estrogen receptorpositive breast cancer histopathology: Comparison to oncotype DX. J Pathol Inform 2011;2:S1.

[28] Pang, B., et al.: Cell nucleus segmentation in color histopathological imagery using convolutional networks. In: CCPR, pp. 15. IEEE (2010)

[29] V. Jain, J.F. Murray, F. Roth, S. Turaga, V. Zhigulin, K.L. Briggman, M.N. Helmstaedter, W. Denk, H. Sebastian Seung. Supervised learning of image restoration with convolutional networks. Proc. ICCV 2007, Rio de Janeiro, Brazil, 2007, vol.1, pp.1-8.

[30] Spyridonos, Panagiota, et al. "Neural network-based segmentation and classification system for automated grading of histologic sections of bladder carcinoma." Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology 24.6 (2002): 317-324.