2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

GENERALIZATION OF PROSTATE CANCER CLASSIFICATION FOR MULTIPLE SITES USING DEEP LEARNING

Ida Arvidsson[⋆]*, Niels Christian Overgaard*[⋆]*, Felicia-Elena Marginean*[†]*, Agnieszka Krzyzanowska*[†]*,*
Anders Bjartell[†]*, Kalle Åström*[⋆]*, Anders Heyden*[⋆]

[⋆] Centre for Mathematical Sciences, Lund University, Sweden
[†] Department of Translational Medicine, Lund University, Sweden

ABSTRACT

Deep learning has the potential to drastically increase the accuracy and efficiency of prostate cancer diagnosis, which would be of uttermost use. Today the diagnosis is determined manually from H&E stained specimens using a light microscope. In this paper several different approaches based on convolutional neural networks for prostate cancer classification are presented and compared, using three different datasets with different origins. The issue that algorithms trained on a certain site might not generalize to other sites, due to for example inevitable stain variations, is highlighted. Two different techniques to overcome this complication are compared; by training the networks using color augmentation and by using digital stain separation. Furthermore, the potential of using an autoencoder to get a more efficient downsampling is investigated, which turned out to be the method giving the best generalization. We achieve accuracies of 95% for classification of benign versus malignant tissue and 81% for Gleason grading for data from the same site as the training data. The corresponding accuracies for images from other sites are in average 88% and 52% respectively.

Index Terms— Convolutional neural network, autoencoder, digital stain separation, prostate cancer, Gleason grade

1. INTRODUCTION

For correct treatment of prostate cancer Gleason grading is of great importance [1]. The grades are determined by pathologists, by ocular inspection of biopsies stained with heamatoxylin and eosin (H&E) in a light microscope, see Figure 1. There are large variations in grading within and between different pathologists [2], and to reduce these as well as speed up the process, an automatic Gleason grading tool would be of great use. As for so many other image analysis tasks deep learning, and especially convolutional neural networks (CNN), has turned out to be very successful [3]. In the last few years multiple papers have presented promising suggestions towards this goal. For example, [4] successfully separated benign and malignant tissue and was able to automatically exclude slides containing only benign tissue. In [5] the false negative rate was reduced to a quarter of a pathologist's for the task of detecting tumors in breast cancer. For Gleason grading, promising results were presented in [6] and [7].

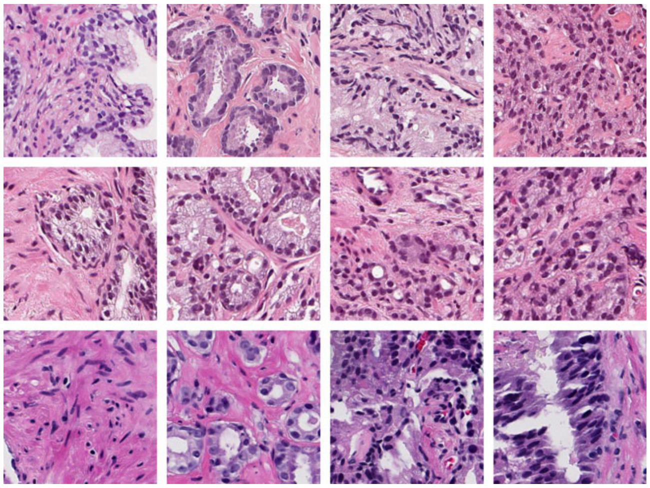

Fig. 1. Examples of H&E stained tissue with different Gleason grades (left to right: Benign, grade 3, grade 4, grade 5) and the different datasets (top to bottom: dataset A, B and C).

In this paper we focus on an issue which has not been much discussed before; algorithms trained to perform well on samples from a certain site might perform substantially worse on samples from other sites. This could be caused by e.g. variations in staining procedures and different equipment. That the algorithms do not generalize was found out by using three datasets with different origins, which have been annotated by the same two pathologists to avoid inter-observer variations.

The purpose of this paper is to investigate the performance of different preprocessing and classification techniques, and their ability of generalization. We use two different CNNs for classification, using images in 20X and 5X magnification respectively. For the latter, three different preprocessing ap-

Thanks to Vinnova (2015-04740) for funding. Athanasios Simoulis helped Felicia-Elena Marginean and Agnieszka Krzyzanowska with scanning and annotating the images. Images from Linköping University Hospital were provided by Claes Lundström. Software support was provided by Sectra.

978-1-5386-3636-7/18/$31.00 ©2018 IEEE

proaches are compared; (i) regular downsampling, (ii) regular downsampling and digital stain separation [8], and (iii) using the encoding part of an autoencoder, trained specifically for this purpose, for downsampling. The aim of using an autoencoder is to get a more efficient downsampling extracting more information than the regular one. Finally, we also investigate to what extent color augmentation of the training data improves the performance and ability to generalize.

2. METHODS

2.1. Data

The prostatic tissue was prepared and stained with H&E and scanned to produce digital images. These were captured in either 40X or 20X magnification, but downsampled to 20X magnification. For this study three different datasets, hereafter referred to as dataset A, dataset B and dataset C respectively, were used. The datasets were created from annotated areas of wholeslide images, where the annotations were made by the same two pathologists at Skåne Univeristy Hospital. The annotations used were ”Benign”, including benign glands, stroma and other benign tissue types, ”Gleason grade 3”, ”Gleason grade 4” and ”Gleason grade 5” [1].

An overview of the datasets is given in Table 1 and examples of their appearances and the different classes are shown in Figure 1. Since the networks required a minimum size of 312 × 312 pixels in 20X magnification, only the annotated areas which were at least this large were included in the datasets. Dataset A, the largest of the three, originates from Skåne University Hospital, Malmö, Sweden. It was used for training and testing of the networks, where three quarters of the annotated areas of each class were used for training and the rest for testing. Dataset B originates from Linköping University Hospital, Linköping, Sweden, and dataset C originates from Erasmus University Medical Center, Rotterdam, the Netherlands. Both these datasets were used for validation, and thus were not used to tune the networks during training.

From the training images patches were created to train the networks. Data augmentation was used to expand the amount of data, using rotation and flipping. The same number of patches was extracted for each class, to not promote any of them. The same patches were used to train all the different networks, although color augmentation was performed for some of the trainings, see Section 2.4 and 3.

The testing and validation was performed on 312 × 312 pixel patches, i.e. for each patch the networks suggested one class only. Thus multiple patches were extracted from each annotated area. Since the classification should be invariant to rotation and flipping each patch was tested 8 times individually; once for each 90° rotation and for their mirrored versions. The same patches were used for testing and validation of all different methods.

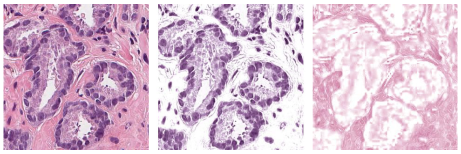

Fig. 2. Example of digital stain separation using the method in [8]: original (left) and the two stain channels (right).

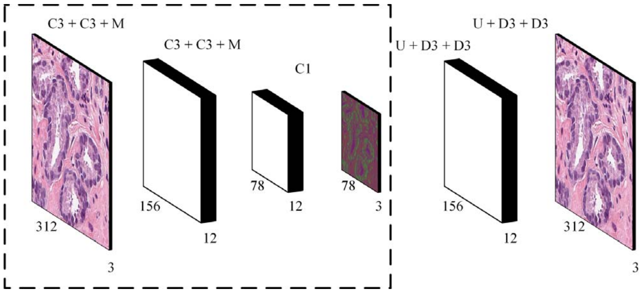

Fig. 3. The design of the autoencoder, where the part in the dashed rectangle is used for downsampling of the data. C3 and C1 means convolution with filters with size 3 and 1, M and U means 2 × 2 max-pooling and up-sampling, and D3 means transposed convolution [9] using filters with size 3.

2.2. Preprocessing Techniques

Two non-standard methods for preprocessing were used; digital stain separation and normalization as well as autoencoders.

The aim of using digital stain separation and normalization was to remove variations in hue and intensity of the stains between different sites. The implementation provided by [8], which is based on sparse non-negative matrix factorization, was used. The output was normalized to $[-1, 1]$. As input to the classifier we used the two stain channels, see Figure 2.

An autoencoder was trained for efficient downsampling of the images instead of using regular downsampling. Since the images should be downsampled from 20X to 5X, an autoencoder with two max-pooling layers and zeropadding for the convolutions was used. The design is shown in Figure 3. All convolutions were followed by a rectified linear unit (ReLU) activation function, except the C1 convolution in the bottom of the network where softmax was used and the last C3 convolution where the hyperbolic tangent was used. The softmax was chosen to try forcing the network to choose one of the channels and thus ideally getting the different channels to represent different types of the tissue. Since the images were normalized to $[-1, 1]$ the hyperbolic tangent was used as the last activation function. The first half, including the C1 layer, was used for downsampling the data. Illustration of a three channel encoded result as an RGB image is shown in Figure 4, and illustrates what was used as input to the classifier.

Table 1. Details of the datasets used for training and testing (A) and validation (B and C).

Dataset	Number of slides	Number of annotated areas				Number of patches per class for training	Number of patches per class for testing and validation
		Benign	G3	G4	G5		
A	88	1527	343	411	113	76320	590
B	15	48	9	171	17	-	600
C	24	24	165	106	3	-	49

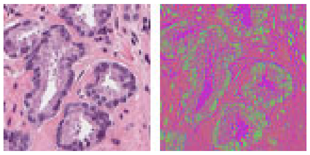

Fig. 4. Example of an input image and corresponding result when the image is encoded with the autoencoder. The intensity of the result is increased, for a more clear visualization.

2.3. Classifiers

The two different CNNs used for classification, illustrated in Figure 5, have the difference that the 20X classifier has two extra layers of convolutions, using zeropadding, and maxpoolings. The 5X network was inspired by the networks presented in [4, 6], using a similar structure and input resolution and no zeropadding. ReLU was used as activation function after each convolution except the last where softmax was used.

2.4. Training of the Networks

The networks were implemented in Keras [10] and trained using the Adam optimizer [11]. When training the classifiers 50% dropout was used between the three last fully connected layers. All input images were, after the preprocessing, normalized to $[-1, 1]$.

In addition to spatial augmentation we also augmented the colors of the images, see Figure 6. The purpose of this is to force the network to be invariant to the inevitable variations in stain appearances and instead learn spatial structures. This method is substantially different from using stain separation and normalization. The color augmentation was constructed by for each image converting the RGB information to the HSV space and randomly and independently change the H, S anv V values in the intervals $[H/1.04, H \cdot 1.04]$, $[S/1.25, S \cdot 1.25]$ and $[V/1.25, V \cdot 1.25]$ respectively, and then transforming back to an RGB image. The autoencoder was trained using color augmentation, while only some of the classifiers were trained with color augmentation, see Section 3.

3. RESULTS

An overview of the accuracies for the different methods as well as the different test and validation datasets is given in Table 2. For all methods the results for benign versus malignant (i.e. Gleason grade 3, 4, 5 together) were superior to the individual Gleason grading results although the classifiers were trained for individual Gleason grades. Furthermore, all methods generalized quite well for benign versus malignant, while the results for Gleason grading were significantly worse for the validation datasets. Since the datasets are annotated by the same pathologists, there ought to be no difference due to inconsistency between pathologists. However, it could be due to inconsistency within these pathologists.

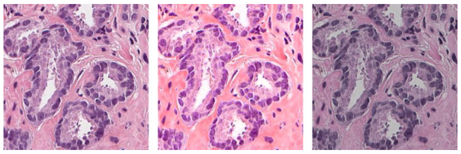

Fig. 6. Illustration of a patch (left) and two of its most extreme color augmentations (right).

Approximately the same accuracies were obtained with the different methods, but the highest for the test set was nevertheless achieved with the network using the highest input resolution, suggesting that some important information was lost when downsampling to 5X. The best results for the validation datasets were obtained when the autoencoder was used for downsampling and the classifier was trained using color augmentation. Thus the autoencoder seems to make a cleverer downsampling than the regular downsampling. The use of color augmentation increased the network's ability to generalize, even when regular downsampling was used. The improvement when using the digital stain separation and normalization was similar to when color augmentation was used.

4. CONCLUSION

We have compared several different methods for prostate cancer classification, and have found out that while the methods do not generalize for the task of Gleason grading they still manage to separate benign and malignant tissue in slides from different sites. Best generalization was obtained when an autoencoder was used for downsampling, while the highest accuracy was achieved with the classifier using the highest resolution of the input images. Furthermore, both color augmentation of the training data as well as the use of digital stain separation and normalization increased the ability to general-

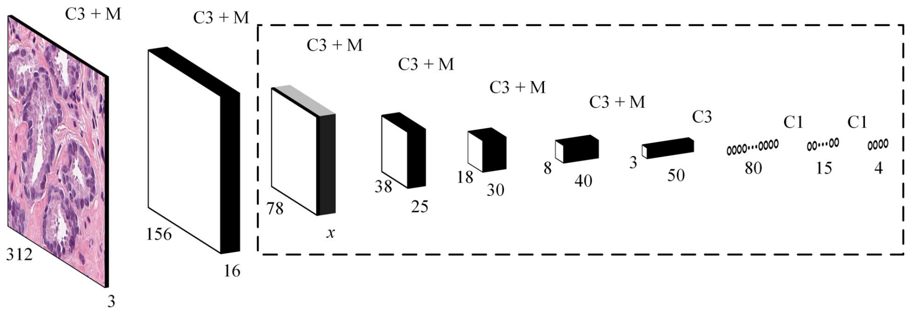

Fig. 5. The design of the CNN classifiers, where inputs in 20X uses the whole network and inputs in 5X uses the part in the dashed rectangle. C3 and C1 means 3×3 and 1×1 convolution, M means 2×2 max-pooling. The number of features x in the third layer depends on the input type; $x = 16$ for 20X, $x = 3$ for 5X using regular downsampling or the encoder and $x = 2$ for digital stain separation.

Table 2. Results on the test and validation datasets, where the first percentage gives the accuracy for Gleason grading and the one in parenthesis the accuracy for benign versus malignant.

Input size &	Test accuracy (%)	Validation accuracy (%)	
preprocessing	Dataset A	Dataset B	Dataset C
20X – color augmentation	**81 (95)**	46 (76)	50 (92)
5X	73 (94)	46 (80)	45 (88)
5X – color augmentation	79 (95)	42 (81)	49 (92)
5X – stain normalization	78 (94)	48 (79)	**53** (89)
5X – autoencoder	75 (92)	50 (83)	47 (90)
5X – autoencoder, color augmentation	75 (93)	**53 (83)**	50 **(94)**

ize. Future work would be to extend the use of autoencoders and explore the possibility of making them site-specific.

5. REFERENCES

[1] J. I. Epstein, M. J. Zelefsky, et al., "A contemporary prostate cancer grading system: a validated alternative to the gleason score," *European urology*, vol. 69, no. 3, pp. 428–435, 2016.

[2] J. Persson, U. Wilderäng, et al., "Interobserver variability in the pathological assessment of radical prostatectomy specimens: findings of the laparoscopic prostatectomy robot open (lappro) study," *Scandinavian journal of urology*, vol. 48, no. 2, pp. 160–167, 2014.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] G. Litjens, C. I. Sánchez, et al., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, pp. 26286, 2016.

[5] Y. Liu, K. Gadepalli, et al., "Detecting cancer metastases on gigapixel pathology images," *arXiv preprint arXiv:1703.02442*, 2017.

[6] A. Gummeson, I. Arvidsson, et al., "Automatic gleason grading of h&e stained microscopic prostate images using deep convolutional neural networks," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2017, pp. 101400S–101400S.

[7] O. Jimenez-del Toroab, M. Atzoria, et al., "Convolutional neural networks for an automatic classification of prostate tissue slides with high–grade gleason score," in *Proc. of SPIE Vol*, 2017, vol. 10140, pp. 101400O–1.

[8] A. Vahadane, T. Peng, et al., "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[9] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[10] F. Chollet et al., "Keras," `https://github.com/fchollet/keras`, 2015.

[11] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.