Convolutional neural networks for an automatic classification of prostate tissue slides with high–grade Gleason score

Oscar Jimenez–del–Toro[a b], Manfredo Atzori[a], Sebastian Otálora[a b], Mats Andersson[c], Kristian Eurén[c], Martin Hedlund[c], Peter Rönnquist[c], and Henning Müller[a b d]

[a]University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland; [b]University of Geneva, Geneva, Switzerland; [c] ContextVision AB, Stockholm, Sweden; [d]Martinos Center for Biomedical Imaging, Charlestown, USA;

ABSTRACT

The Gleason grading system was developed for assessing prostate histopathology slides. It is correlated to the outcome and incidence of relapse in prostate cancer. Although this grading is part of a standard protocol performed by pathologists, visual inspection of whole slide images (WSIs) has an inherent subjectivity when evaluated by different pathologists. Computer aided pathology has been proposed to generate an objective and reproducible assessment that can help pathologists in their evaluation of new tissue samples. Deep convolutional neural networks are a promising approach for the automatic classification of histopathology images and can hierarchically learn subtle visual features from the data. However, a large number of manual annotations from pathologists are commonly required to obtain sufficient statistical generalization when training new models that can evaluate the daily generated large amounts of pathology data. A fully automatic approach that detects prostatectomy WSIs with high–grade Gleason score is proposed. We evaluate the performance of various deep learning architectures training them with patches extracted from automatically generated regions–of–interest rather than from manually segmented ones. Relevant parameters for training the deep learning model such as size and number of patches as well as the inclusion or not of data augmentation are compared between the tested deep learning architectures. 235 prostate tissue WSIs with their pathology report from the publicly available TCGA data set were used. An accuracy of 78% was obtained in a balanced set of 46 unseen test images with different Gleason grades in a 2–class decision: high vs. low Gleason grade. Grades 7–8, which represent the boundary decision of the proposed task, were particularly well classified. The method is scalable to larger data sets with straightforward re–training of the model to include data from multiple sources, scanners and acquisition techniques. Automatically generated heatmaps for the WSIs could be useful for improving the selection of patches when training networks for big data sets and to guide the visual inspection of these images.

Keywords: Prostate cancer grading, Convolutional neural networks, Computer aided Pathology

1. INTRODUCTION

Prostate cancer is the second most common cancer in men.[1] The Gleason grading system is widely accepted as a part of a standard protocol when assessing prostate adenocarcinoma (PRAD) histopathological samples (biopsies or prostatectomies).[2] It is correlated to the prognostic factor, patient outcome, local recurrence and risk of metastasis.[3] The grading is based on the architectural patterns shown in prostate tissue samples that describe tumor appearance and the presence of alterations in the glands.[4] The final Gleason grade results from the sum of the two patterns (from 1 to 5) most present in the tissue slide producing a final grade in the range of 2 to 10. Particularly, in a Gleason grade $\geq$ 8, cells are poorly differentiated, prognostic is adverse with a higher risk of relapse and the 5–year biochemical risk–free survival ocurrs in less than half of these cases (48%).[5] Therefore, it is important in clinical practice to objectively differentiate between lower and higher Gleason scores for the treatment and follow up of the patients. However, the pathologists grading can vary according to their personal experience, volume of practice and inherent subjectivity.[6]

Computer aided pathology has been proposed in the past years as an option for generating an objective and reproducible score that could help pathologists in their evaluation of new tissue samples.[7] With the advent of digital pathology, mostly due to feasibility of high–resolution whole slide imaging (WSI), multiple approaches

Further author information: Send correspondence to O.J.T. Email: oscar.jimenez@hevs.ch have been proposed to improve the detection and classification of the slides compared to only performing visual inspection.[8] Some of the proposed approaches in literature have used hand–crafted morphological and architectural features to classify tumors according to their phenotype differentiation.[9,10] In recent years, deep learning (DL) has been a successful machine learning paradigm for various fields such as pattern recognition in computer vision[11,12] and has also been recently proposed for some digital pathology tasks[7,13,14] and also combinations of deep learning based and other features have been proposed.[15] Most of the progress in the performance of deep learning approaches (i.e. convolutional neural networks) can be attributed to larger training data sets being available that improve statistical generalization, and powerful computers with better software infrastructure that were non–existent until a few years ago.[16]

Convolutional neural networks (CNNs) are a kind of neural networks that use a linear operation called convolution in at least one of their layers to process input data with a grid–like topology (e.g. image data).[17] As a fundamental property, CNNs assume shift–invariance i.e. a pattern is detected the same way wherever it appears in the image, which makes them very efficient. CNNs learn hierarchical representations of the data with nonlinear composition of features from the first layers up to the the highly semantic ones present in the later layers near where the classification is performed. This technique has often obtained better results when compared to many other texture descriptors that are not able to represent the subtle patterns embedded in biological tissues so precisely.[18] However, the training of these models for histopathology image classification is generally based on extracting patches from manually annotated images with delineated regions–of–interest containing tumor and normal areas in the WSI. Manually annotating tumors in WSI is time consuming and not scalable to the large number of slides that are produced daily in hospitals.[14] Moreover, rough manual annotations can lead to missed isolated tumor cells[14] and imprecise borders of the segmented tumors.[19] In this paper we propose an approach that automatically generates patches from WSIs in regions–of–interest (so from the tumor areas) and trains a CNN model to classify images with high Gleason grades (≥ 8) using a large data set of non–manually segmented tissue slides. We explore the possibility to train a model using only the sum of the WSI Gleason score if the samples are selected to represent high Gleason grades from the images.

2. METHODS

2.1. Data set

The Cancer Genome Atlas (TCGA) Research Network established a bio–specimen repository including a large collection of digital pathology whole slide images from various patient tissues[20][*]. These tissue samples were collected from multiple sites during clinical practice and from different scanners, manufacturers and acquisition protocols. The TCGA images are regarded as 'tissue slides', geographically adjacent to the tissue samples used as 'diagnostic slides' on which the reports are based. For a subset of the images, the pathology reports from the time of the sample acquisition, are also available. This subset contains around 400 pathology images with their respective pathology reports. For this paper, 235 PRAD Hematoxylin and Eosin (H&E) stained WSIs of radical prostatectomies were used together with the grade extracted by a physician from their corresponding pathology reports (a sample image is shown in Fig. 1 A). We make this TCGA subset (as well as their extracted grades) publicly available for future research on this data set[†]. The Gleason patterns and final grade for each slide was taken from the provided pathology reports. In cases where the report mentioned more than one Gleason grade during the treatment of the patient, the final grade given to the 'diagnostic slide' was selected. As it is common in histopathological slides, some images can contain multiple samples in a single image, in those cases all samples were included during the experiments. All images were digitized with a 40× objective.

The data set was randomly divided: 60% for training (141 images), 20% for validation (47 images) and 20% for test (46 images). The test WSIs were not used during the training or the optimization of the CNN model. The PRAD cases were selected randomly from those available in the TCGA data set where the lowest Gleason grade available in the images was 6 and the highest grade was 10. The number of images per Gleason grade was the following: G6, 19; G7, 87; G8, 39; G9, 83; G10, 7. For the experimental set up, we aimed at dividing the data set into images with lower Gleason (grade ≤ 7) score and higher Gleason (grade ≥ 8) score. This resulted

[*] `http://cancergenome.nih.gov/`, as of 1 January 2017

[†] `https://goo.gl/BBzsn6`

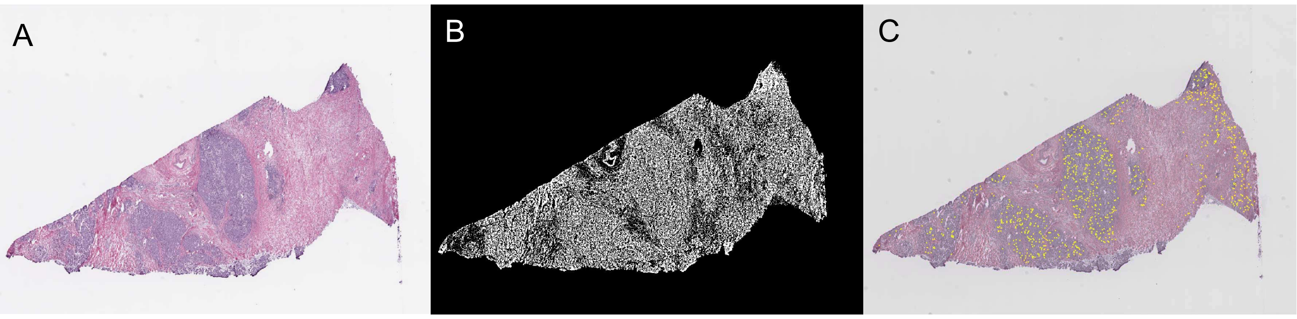


Figure 1: Whole-slide sampling. **A**: Normal H&E tissue slide. **B**: Tissue mask generated using the Blue Ratio image. **C**: Selected patches (in yellow) from the 40× resolution of the image used for training.

in an evenly divided data set of grades 6 and 7 with 106 images (45%) and grades 8, 9 and 10 with 129 images (55%).

2.2. Selecting relevant patches for Gleason grading

As with the majority of WSI data sets, the images obtained from the TCGA data set do not contain any manual annotation of regions–of–interest on which the pathologists have based their grading. To train a model that classifies WSI we first detect the areas with the most relevant visual features for a posterior patch extraction. A binary tissue mask (Fig. 1 B), that can remove more than 80% of empty image content, is generated using the Blue Ratio image (BR) in the 5× resolution of the image. BR images are commonly used in digital pathology to detect the nuclei from the cells because these structures appear blue in (H&E) stained images, whereas the extracellular material appears more pink or red.[21] The BR image is computed as follows:

$$BR = \frac{100 \times B}{1 + R + G} \times \frac{256}{1 + B + R + G} \tag{1}$$

where B, R and G are the blue, red and green channel of RGB, respectively. The first part maps the hue and the second part maps the intensity. The mask is then slightly dilated and closed with morphological operations to have a mask of the tissue in the WSI.

Evaluating the full image can be computationally expensive and including many samples from non–affected tissue areas can hamper the overall Gleason grading classification. Randomly generated 256 × 256 patches are extracted from the tissue mask. Different number of patches were tested during optimization of the algorithm. For the final evaluation of the algorithm, a total of 3000 patches per image was selected. These patches, originally in the RGB space, are then transformed to their BR image with the same equation as Eq. 1. The aim is to select those with the highest energy in the BR space (Fig. 2) and better represent the diversity of tumor areas without including many patches from normal areas. These patches correlate well to the tumor regions in the image and are uniformly distributed so that relevant areas are well represented when training a CNN.

2.3. Training a CNN to classify high–grade Gleason scores

The open source deep learning framework Caffe was used to train and evaluate the convolutional neural networks.[22] Different deep neural network architectures were tested during the optimization of the algorithm: LeNet,[17] AlexNet[11] and GoogleNet.[12] This networks have been shown to improve natural image classification systematically using labeled training data sets of size in the order of million of samples. The theory of domain adaptation and transfer learning have shown that reusing the parameters learnt in one particular dataset or domain on another one could alleviate the necessity of a large manually annotated dataset.[23] Particularly, reusing existing networks significantly reduces the convergence time required to find the most successful parameters of the network in the new domain. The tested networks have been previously evaluated in open benchmarks and new upgrades could be added in a straightforward manner to the proposed approach.

Each architecture was trained using the same patches obtained from the previous patch selection step with 40× resolution and testing various input sizes and data augmentation techniques. The patches were mean

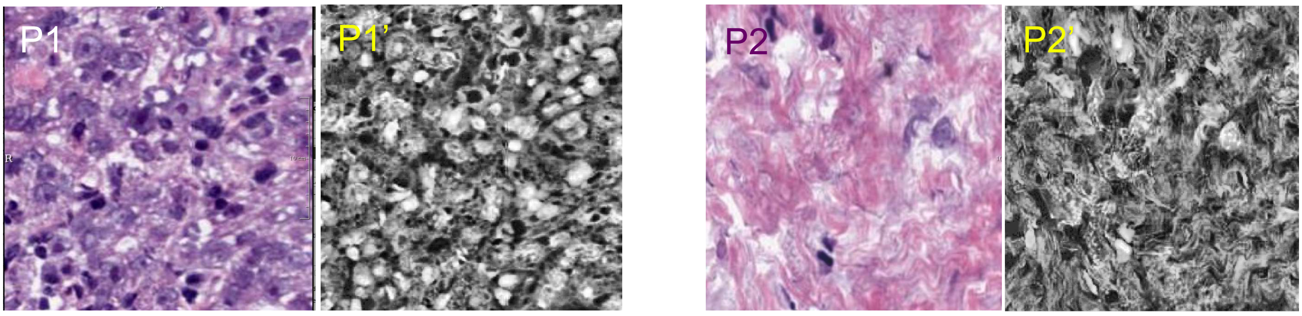


Figure 2: Patch selection. Two patch samples (**P1**,**P2**) with their corresponding BR image (**P1'**,**P2'**). Patch 1 is one of the top patches selected to represent the regions–of–interest from the WSI. Patch 2 is in the bottom of the ranking according to the blue ratio, therefore it is not included in the final selection of patches for the training. Notice how a higher number of cell nuclei increases the energy of the blue ratio and better represents the areas on which a Gleason score should be based.

corrected and fed to the networks in their original RGB color space, we also tested in gray scale for the network with the best classification accuracy. As mentioned in the data set description, both classes (high vs low Gleason grade) had a fairly similar number of images and consequently a similar number of patches during training. In total, aprox. 150k patches (without considering data augmentation) were included in the train set for class Gleason 6–7 (low Gleason) and aprox. 150k for class Gleason 8–10 (high Gleason). Patch batches with a fixed size of 64 were iteratively fed to the network over a series of epochs and the optimal parameters were estimated through stochastic gradient descent optimization. Softmax was used as a loss function in the last layer and the training was stopped if there was no significant improvement after 5 epochs in the overall model accuracy. Once the highest accuracy was achieved, the learned weights are no longer modified and patches from unseen images are classified with the resulting model. The optimal performance for the various networks evaluated was obtained in around 5–10 epochs. After training the networks with the train and validation set, the models were used on the test set without any further optimization of the parameters.

3. EXPERIMENTAL RESULTS

The following parameters were evaluated to define the best model for correctly classifying patches as lower or higher Gleason grade:

- Size of the patches: 128 × 128, 256 × 256
- Number of input patches per slide: 1000, 2000
- Channels: RGB, grayscale
- Type of deep learning architecture: LeNet, AlexNet, GoogLeNet
- Using data augmentation or not: mirroring and random cropping of sub–images

The results for each combination of parameters is shown in 1. The model that gave the highest accuracy was GoogLeNet (27 layers deep) with an input of 2000 RGB patches per image, size 256 × 256 and with data augmentation (mirroring and randomly cropping to images of size 227 × 227). The model took 1 day to train using 2 NVIDIA Tesla K80 GPUs, and was stopped after 12 epochs because of little or no improvement in the accuracy.

In order to classify whole slides from the selected patches per image, an optimization was performed on the percentage of patches with higher grading per slide to be considered as a high Gleason grade slide. With a percentage of at least 40% of the patches classified as high grading, 36 out of 46 (78.2% accuracy) WSIs were correctly classified in the test set. The best classified scores were images with Gleason grade 10 (1 image) and

Table 1: Testing different parameters for training a neural network to classify lower vs. higher Gleason grade patches. The 'Patch size' is the original size in pixels of the patches in 40× resolution. 'Num. patches' is the number of patches selected within the tissue and classified with higher probability of being inside tumor areas. 'Channels' is the color space of the patches fed to the network: either RGB or 'grayS' (RGB patches transformed to gray–scale intensity). 'Data augment.' highlights if there was data augmentation: mirroring and randomly cropping a smaller sub–image inside the original patch. 'Accuracy' is the result of the trained model in the total patches from the validation set.

Architecture	Patch size	Num. patches	Channels	Data augment.	Accuracy
LeNet	128	2000	RGB	Yes	63.85
AlexNet	128	2000	RGB	Yes	64.55
GoogLeNet	128	2000	RGB	Yes	69.58
LeNet	227	2000	RGB	Yes	69.47
AlexNet	227	2000	RGB	Yes	71.16
GoogLeNet	**227**	**2000**	**RGB**	**Yes**	**73.52**
GoogLeNet	227	2000	grayS	Yes	69.12
GoogLeNet	227	2000	RGB	No	70.37
GoogLeNet	227	1000	RGB	Yes	70.38

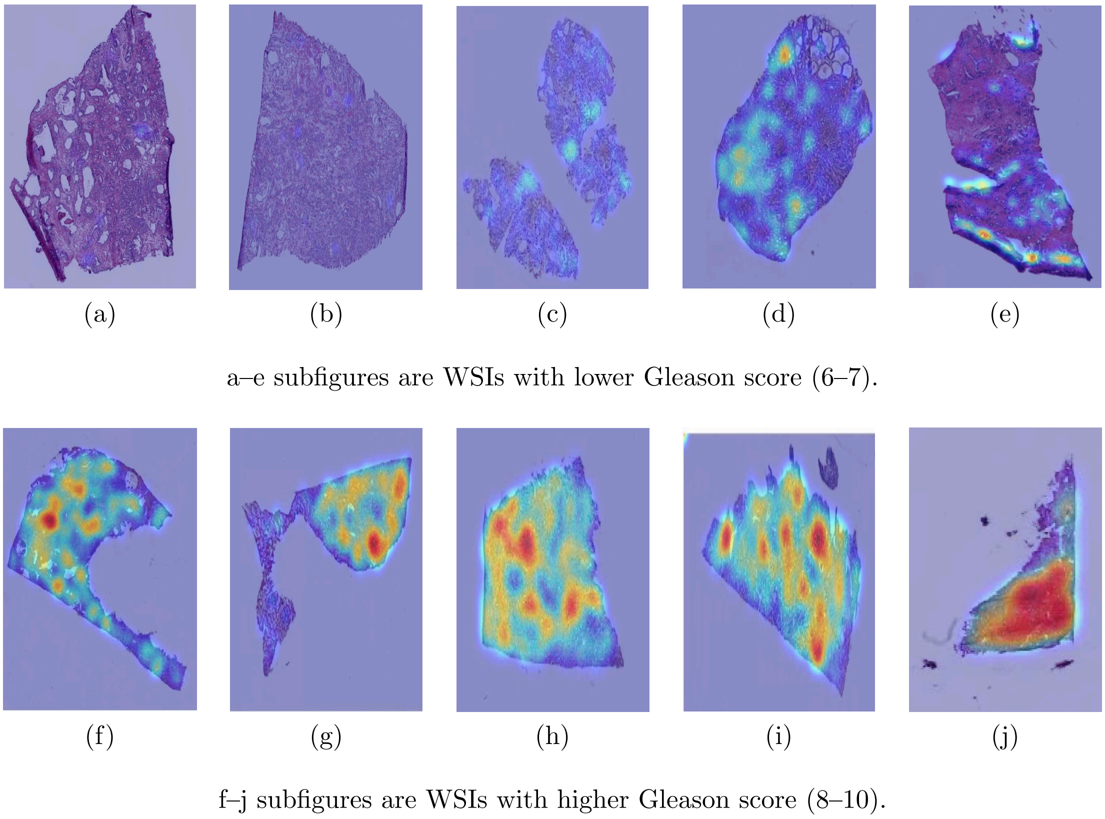
Gleason grade 8 (8 images) with 1.00 and 0.862 accuracy respectively. The images with the worst Gleason grade classification were images with grade 6 with only 1 out of 4 images classified correctly with a grade ≤ 7.

To qualitatively inspect the results from the classification, heatmaps were created using the interpolated patch coordinates on the 5× resolution. The probability from each patch sample of being representative from high Gleason grades was overlaid on the whole slide images in a scale of 0 to 255, with blue color representing a low probability and red representing a high probability. The heatmaps were smoothed using a σ of 2 to generate a smoother visualization. A few sample heatmaps are shown in Fig. 3. For a closer view of the heatmap overlay at a high power field (40×), please refer to Fig. 4.

4. DISCUSSION

A method for detecting whole slide images from radical prostatectomies with high Gleason grades is proposed. The approach was tested on a large publicly available data set and obtained an overall accuracy of 78% in a 2–class setup: low (6–7) vs high (8–10) Gleason grades. The method is fully automatic and able to select patches from regions–of–interest from non manually segmented images for training a deep convolutional neural network and classify a WSI in approximately 5 minutes. Although images with the lower grade (6) were the worst classified (only 4 cases in the test set), the images in the selected Gleason grading threshold (7–8) were mostly correctly classified, meaning that with a larger number of samples for every grade the results could potentially still be improved. This method could be used by pathologists to reduce the subjectivity of visually exploring large WSIs and reduce the interobserver variability that still occurs when pathology reports are made.

In Fig. 3 different tissue samples from the test set are shown with the generated heatmaps overlaid on the 5× resolution image. In subfigures a–e images with lower Gleason grades show fewer areas with activation for high Gleason patterns. Additionally, in the highlighted areas, the heatmaps only show a mild activation, except for subfigure (e) where a few areas are highlited in foldings of the tissue sample. These type of foldings and tearing of the tissue are common in pathology studies and should be easy to detect for a pathologist. Nevertheless, a more in depth evaluation of the impact of tissue foldings in the classification is out of the focus of this work. On the other hand in Fig. 5, a single whole slide image with multiple tissue samples is shown. These tissue samples were adjacent to each other in the prostate and should share similar visual patterns. Even though the approach is the same as for images with a larger single tissue sample, all the samples show highlighted areas in the same spatial location. This example is presented to emphasize the robustness of the method in the localization of areas with a higher probability of containing higher Gleason grade patterns. The BR intensity amplifies the areas with high nuclei density within a WSI, still it is hard to directly compare the BR values between different

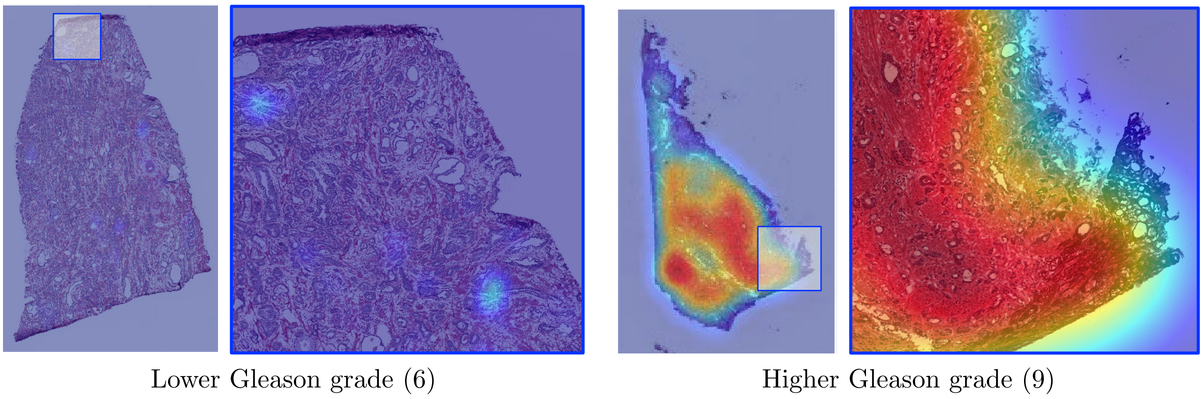

(a) (b) (c) (d) (e)

a–e subfigures are WSIs with lower Gleason score (6–7).

(f) (g) (h) (i) (j)

f–j subfigures are WSIs with higher Gleason score (8–10).

Figure 3: Prostatectomy whole slide images of the test set with overlaid heatmaps of the probability of samples being from a higher Gleason grade. The heatmaps were generated with the classification of test set patches using the deep learning model that produced the higher accuracy in the validation set without any further optimizations. Areas with blue color represent a none or lower probability of higher Gleason grade, whereas red areas highlight a higher probability of higher Gleason grade scores. The selected tissues samples were individually cropped from the whole slide images to facilitate the visualization of the heatmaps.


Lower Gleason grade (6)

Higher Gleason grade (9)

Figure 4: Examples of a lower Gleason grade (6) WSI and a higher Gleason grade (9) WSI heatmaps with their corresponding magnification at 40× resolution (in the highlighted blue square of the images on the right). The heatmaps were computed for the WSI at 5× resolution and zoomed in for their visualization at a high power field. More dense sampling can be performed at the higher resolutions in regions–of–interest to have a smoother representation of the heatmaps at this scale.

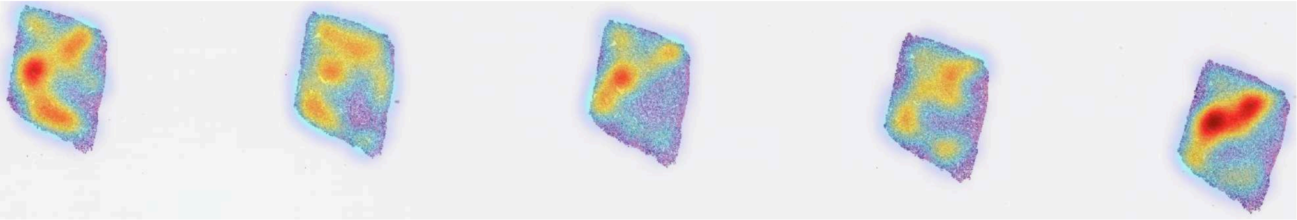


Figure 5: Heatmap overlaid on a single whole slide image (Gleason grade 8) with four adjacent tissue samples. Although the samples were selected in the same fashion as WSIs with a single tissue sample, the four samples show similar areas with high probability of a higher Gleason grade (≥ 8). This shows the robustness of the method for spatially localizing areas with higher probability of higher Gleason scores.

WSIs due to variations in the staining. Staining normalization techniques and more complex approaches for tumor segmentation have been proposed in the literature.[8, 24] Their integration into the proposed method is straightforward as well as the selection of an alternative architecture for the deep learning classification. Although the task was set as a 2–class classification problem, future work should address the feasibility of ranking every WSI to their corresponding Gleason grade. The limitation into two classes is convenient for a first experiment and it is straight forward the extend the proposed concept to more classes. The aim of this work was to evaluate the feasibility of using automatically generated samples from non manually segmented WSIs to train a reliable model for Gleason grade classification. The motivation for this approach is that manual annotations are not scalable to the amount of images generated daily in hospitals as this is a time consuming task particularly for these very large images. A limited amount of images with rough manual annotations are frequently used to train these type of networks, however these automatically generated heatmaps could potentially reduce the time spent in manual annotations by already highlighting regions–of–interest. Moreover, they could potentially be used to include a larger amount of images to train deep learning networks than those that could be manually annotated. Targeted manual annotations can also potentially improve results with a limited effort with approaches such as active learning.

To train and objectively test image analysis systems for diagnostic aid, large annotated data sets (i.e. TCGA) are required.[25] Open competitions targeting common tasks in the image analysis of pathology images have been proposed in recent years.[8] The advantage of having different methods tested on the same tasks with a common data set is that the strengths and limitations of state–of–the–art approaches can be objectively compared. Particularly in the case of histopathology, whole slide image classification has been addressed in CAMELYON16[‡] for detecting cancer metastasis in lymph nodes and TUPAC16[§] for assessing tumor proliferation in breast images. Our approach is similar to the winning approach from CAMELYON16 of Wang et al.[26] , nevertheless in this paper an automatic selection of patches is performed without having manual image segmentations available. This supports our hypothesis that is also feasible to generate large data sets of annotated pathology slides automatically. Selecting the most suitable approach to support and speed–up the visual interpretation of the images could help pathologists focus on the more pressing issues when diagnosing these studies.

5. CONCLUSIONS

A fully automatic approach using patch selection and CNN to classify whole slide prostate images is presented in this paper. The approach detects regions–of–interest in WSIs where relevant visual information can be sampled to detect high–grade Gleason scores. No manual segmentations were used for the training of these networks. Nevertheless, the results are promising for the classification and visualization of pathology tissue samples. Visual heatmaps of the areas with higher probability of presenting high Gleason patterns were generated and show consistent spatial representations of the data. The approach is scalable to large data sets and can generate faster results once a model is trained with multiple images. More objective and repeatable assessment of these images

[‡] `https://camelyon16.grand-challenge.org`, *

[§] `http://tupac.tue-image.nl`, * as of 10 January 2017 (using methods similar to the one proposed in this paper) could be helpful for pathologists to reduce inter– and intra–observer variability.

ACKNOWLEDGMENTS

This work was partially supported by the Eurostars project E! 9653 SLDESUTO-BOX.

REFERENCES

1. Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer* **136**(5), E359–E386 (2015).
2. Gleason, D. F., "Classification of prostatic carcinomas.," *Cancer chemotherapy reports. Part 1* **50**(3), 125–128 (1966).
3. Humphrey, P. A., "Gleason grading and prognostic factors in carcinoma of the prostate," *Modern Pathology* **17**(3), 292–306 (2004).
4. Epstein, J. I., Zelefsky, M. J., Sjoberg, D. D., Nelson, J. B., Egevad, L., Magi-Galluzzi, C., Vickers, A. J., Parwani, A. V., Reuter, V. E., Fine, S. W., Eastham, J. A., Wiklund, P., Han, M., Reddy, C. A., Ciezki, J. P., Nyberg, T., and Klein, E. A., "A contemporary prostate cancer grading system: a validated alternative to the gleason score," *European Urology* **69**(3), 428–435 (2016).
5. Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., Humphrey, P. A., and the Grading Committee, "The 2014 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system," *The American journal of surgical pathology* **40**(2), 244–252 (2016).
6. Trpkov, K., [*Contemporary Gleason Grading System*], 13–32, Springer New York, New York, NY (2015).
7. Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F., and Madabhushi, A., "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *Journal of Medical Imaging* **1**(3), 034003–034003 (2014).
8. Jimenez-del-Toro, O., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rousson, M., Müller, H., and Atzori, M., "Analysis of histopathology images: From traditional machine learning to deep learning," in [*Biomedical Texture Analysis: Fundamentals, Applications, Tools, and Challenges Ahead*], Depeursinge, A., Al-Kadi, O. S., and Mitchell, J. R., eds., Elsevier (2017).
9. Tabesh, A., Teverovskiy, M., Pang, H.-Y., Kumar, V. P., Verbel, D., Kotsianti, A., and Saidi, O., "Multifeature prostate cancer diagnosis and gleason grading of histological images," *IEEE transactions on Medical Imaging* **26**(10), 1366–1378 (2007).
10. Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., and Tomaszeweski, J., "Automated grading of prostate cancer using architectural and textural image features," in [*2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*], 1284–1287, IEEE (2007).
11. Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," in [*Advances in Neural Information Processing Systems 25*], Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., eds., 1097–1105, Curran Associates, Inc. (2012).
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 1–9 (2015).
13. Källén, H., Molin, J., Heyden, A., Lundström, C., and Åström, "Towards grading gleason score using generically trained deep convolutional neural networks," in [*Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*], 1163–1167, IEEE (2016).
14. Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-van de Kaa, C., Bult, P., van Ginneken, B., and van der Laak, J., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports* **6**, 26286 (2016).

15. Otálora, S., Cruz-Roa, A., Arevalo, J., Atzori, M., Mandabhushi, A., Judkins, A., González, F., Müller, H., and Depeursinge, A., "Combining unsupervised feature learning and riesz wavelets for histopathology image representation: Application to identifying anaplastic medulloblastoma," in [*Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015*], Navab, N., Hornegger, J., Wells, W. M., and Frangi, A., eds., *Lecture Notes in Computer Science* **9349**, 581–588, Springer International Publishing (Oct 2015).
16. LeCun, Y., Bengio, Y., and Hinton, G., "Deep learning," *Nature* **521**(7553), 436–444 (2015).
17. Le Cun, Y., Jackel, L. D., Boser, B. E., Denker, J. S., Graf, H.-P., Guyon, I., Henderson, D., Howard, R. E., and Hubbard, W., "Handwritten digit recognition: Applications of neural network chips and automatic learning," *IEEE Communications Magazine* **27**(11), 41–46 (1989).
18. Janowczyk, A. and Madabhushi, A., "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics* **7** (2016).
19. Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., and Madabhushi, A., "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," in [*SPIE medical imaging*], 904103–904103, International Society for Optics and Photonics (2014).
20. Gutman, D. A., Cobb, J., Somanna, D., Park, Y., Wang, F., Kurc, T., Saltz, J. H., Brat, D. J., Cooper, L. A. D., and Kong, J., "Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data," *Journal of the American Medical Informatics Association* **20**(6), 1091–1098 (2013).
21. Chang, H., Loss, L. A., and Parvin, B., "Nuclear segmentation in H&E sections via multi-reference graph cut (MRGC)," in [*International Symposium Biomedical Imaging*], (2012).
22. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S., and Darrell, T., "Caffe: Convolutional architecture for fast feature embedding," in [*Proceedings of the ACM International Conference on Multimedia, MM'14*], 675–678 (2014).
23. Rozantsev, A., Salzmann, M., and Fua, P., "Beyond sharing weights for deep domain adaptation," *CoRR* **abs/1603.06432** (2016).
24. Leo, P., Lee, G., Shih, N. N. C., Elliott, R., Feldman, M. D., and Madabhushi, A., "Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images," *Journal of Medical Imaging* **3**(4), 047502–047502 (2016).
25. Jimenez-del-Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A. A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., Kontokotsios, G., Langs, G., Menze, B., Salas Fernandez, T., Schaer, R., Walleyo, A., Weber, M.-A., Dicente Cid, Y., Gass, T., Heinrich, M., Jia, F., Kahl, F., Kechichian, R., Mai, D., Spanier, A. B., Vincent, G., Wang, C., Wyeth, D., and Hanbury, A., "Cloud–based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL Anatomy Benchmarks," *IEEE Transactions on Medical Imaging* **35**(11), 2459–2475 (2016).
26. Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H., "Deep learning for identifying metastatic breast cancer," *ArXiv eprints* (Jun 2016).