MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

Automated Detection and Grading of Non—Muscle-Invasive Urothelial Cell Carcinoma of the Bladder

Ilaria Jansen,^{*,†} Marit Lucas,^{*} Judith Bosschier,[‡] Onno J. de Boer,[§] Sybren L. Meijer,[§] Ton G. van Leeuwen,^{*} Henk A. Marquering,^{*,¶} Jakko A. Nieuwenhuijzen,[‡] Daniel M. de Bruin,^{*,†} and C. Dilara Savci-Heijink[§]

From the Departments of Biomedical Engineering and Physics,^{*} Urology,[†] Pathology,[‡] and Radiology and Nuclear Medicine,[¶] Amsterdam UMC, University of Amsterdam, Amsterdam; and the Department of Urology,[§] Amsterdam UMC, VU University Medical Center, Amsterdam, the Netherlands

Accepted for publication
March 5, 2020.

Address correspondence to
Ilaria Jansen, M.D.,
Meibergdreef 9, 1105AZ
Amsterdam, the Netherlands.
E-mail: i.jansen@amsterdamumc.nl.

Accurate grading of non—muscle-invasive urothelial cell carcinoma is of major importance; however, high interobserver variability exists. A fully automated detection and grading network based on deep learning is proposed to enhance reproducibility. A total of 328 transurethral resection specimens from 232 patients were included, and a consensus reading by three specialized pathologists was used. The slides were digitized, and the urothelium was annotated by expert observers. The U-Net—based segmentation network was trained to automatically detect urothelium. This detection was used as input for the classification network. The classification network aimed to grade the tumors according to the World Health Organization grading system adopted in 2004. The automated grading was compared with the consensus and individual grading. The segmentation network resulted in an accurate detection of urothelium. The automated grading shows moderate agreement ($\kappa = 0.48 \pm 0.14$ SEM) with the consensus reading. The agreement among pathologists ranges between fair ($\kappa = 0.35 \pm 0.13$ SEM and $\kappa = 0.38 \pm 0.11$ SEM) and moderate ($\kappa = 0.52 \pm 0.13$ SEM). The automated classification correctly graded 76% of the low-grade cancers and 71% of the high-grade cancers according to the consensus reading. These results indicate that deep learning can be used for the fully automated detection and grading of urothelial cell carcinoma. (*Am J Pathol* 2020; 190: 1483–1490; <https://doi.org/10.1016/j.ajpath.2020.03.013>)

Urothelial cell carcinoma (UCC) is the most common type of bladder cancer and is a major challenge in urologic oncology due to its propensity to recur and progress.¹ According to the European Organization for Research and Treatment of Cancer, the histologic grade of UCC is one of the most important prognostic factors for the prediction of recurrence and progression, along with the number and size of the tumor(s), prior recurrence rate, and concomitant presence of carcinoma *in situ*.²

In 1973, the World Health Organization (WHO) introduced a grading system dividing the histologic spectrum of UCC into three grades (WHO'73).^{3,4} Grade 1 represents the lowest grade of cytonuclear atypia, and grade 3 has the highest grade of atypia. Despite the worldwide use of this grading system, the interobserver agreement ranges from

38% to 89%.^{5,6} Moreover, grade 2 is considered a poorly defined category (defined as between grade 1 and grade 3) with widely varying clinical behavior and prognosis.^{7–9} For these reasons, the International Society of Urologic Pathologists proposed a revised classification system that was adopted by the WHO in 2004 (WHO'04).¹⁰ This system divides the atypical spectrum into papillary urothelial neoplasm of low malignant potential, low-grade papillary urothelial carcinoma, and high-grade papillary urothelial carcinoma.¹¹ Papillary urothelial neoplasm of low malignant

Supported by ITEA3 grant ITEA151003 (H.M.).

I.J., M.L., D.M.d.B., and C.D.S.-H. contributed equally to this work.

Disclosures: H.A.M. is founder and shareholder of Nico-lab; D.M.d.B. is founder and shareholder of Offroad Medical.

potential was in the WHO'73 grading system classified as grade 1, whereas low-grade papillary urothelial carcinoma contains both grade 1 and the lower spectrum of grade 2. High-grade papillary urothelial carcinoma contains the higher spectrum of grade 2 and grade 3. Grading according to the WHO'04 guidelines shows higher agreement among pathologists, ranging from 43% to 100%. Nonetheless, the clinical value has not been sufficiently validated against the WHO'73 guidelines.¹ Furthermore, the most commonly used risk assessment tools for the prediction of recurrence and progression are still based on the WHO'73 grading system.^{2,12} Because treatment selection and risk stratification tools are based on either the WHO'73 or the WHO'04 system, the use of both grading systems is advised.⁹

To aid clinical decision-making and treatment selection, a reliable and reproducible methodology is needed for the classification of transurethral resection of bladder tumor (TURBT) specimens. Only a few studies for the automatic grading of UCC have been reported.^{13,14} The existing methodologies use preselected regions of interest, often combined with the use of immunohistochemistry staining.¹⁴ These studies used a combination of hand-crafted textural and morphologic features to differentiate between the WHO'73 grades. However, this approach limits the generalizability and use of the designed methodologies because it still requires human input to select the region of interest, as well as for the feature extraction, and is therefore not fully reproducible. However, the use of state-of-the-art deep learning methodologies in histopathology images has proven to be successful in the detection and grading of various tumor types.^{15–17} Hence, in the current study, we investigated the feasibility of a fully automated detection and grading of TURBT specimens. Two individual neural network architectures were subsequently trained and validated. The first network focused on the detection and segmentation of urothelium, which was used as input for the second network, to grade the digitized sections. Finally, the automated grading was compared with the grading of three experienced pathologists (C.D.S.-H. and two non-author pathologists).

Materials and Methods

Patient Selection

The Institutional Review Board of the Amsterdam UMC (Academic Medical Center, Amsterdam, the Netherlands) granted approval for this study (W18_056 #18.074). A total of 328 non-muscle-invasive bladder cancer (NMIBC) specimens from 232 patients from three different centers, namely Amsterdam UMC (location VUmc, Amsterdam, the Netherlands), Amstelland Hospital (Amstelveen, the Netherlands), and Amsterdam UMC (location Academic Medical Center, Amsterdam, the Netherlands), were included. All patients underwent a TURBT procedure between February 2000 and August 2016. These specimens were obtained within the scope of the study of Bosschieter

et al¹⁸ with the aim to assess reproducibility and prognostic performance of the WHO'73 and WHO'04 guidelines for patients with NMIBC. Hematoxylin and eosin–stained slides from 4 µm thick sections from resection materials were digitized by using the Philips IntelliSite Ultra Fast Scanner (Philips Digital Pathology Solutions, Best, the Netherlands). The digitized sections were exported with 20× resolution, resulting in an in-plane resolution of 0.5 µm per pixel.

Annotations and Grading of Tumors

All exported images were annotated by using a free-hand annotation tool developed in-house.¹⁹ All regions containing UCC were delineated, as well as a selection of regions with non-atypical urothelium and large regions of fibrovascular tissue. Stalks of fibrovascular tissue are often present in tumorous tissue and can vary greatly in size depending on the section plane. All annotations were made by expert observers (residents) and then checked by a specialized uropathologist (C.D.S.-H.). Regions with tissue folds, mechanical damage such as cauterization artifacts, loose regions due to TURBT extraction or sectioning, or regions that are out of focus were annotated as nondiagnostic because the pathologists could not assess the grade on those regions. However, these regions were not meticulously annotated. A representative example of the delineations is shown in Figure 1.

As part of an earlier study,¹⁸ all cases had been previously graded by three experienced pathologists with uropathology as the major field of interest. Briefly, the grade of the tumors was initially assessed according to the worst pattern using the WHO'73 and WHO'04 grading systems, followed by a subsequent consensus reading in case of disagreement. Several slides were available per patient. However, it was not known on which slide the pathologists based the grading.

Automated Classification of Urothelium

To grade the tumor, a two-step approach was used. First, the U-Net–based segmentation network was trained to automatically detect urothelium in the sections. The second step involved the setup of a classification network, to automatically grade the segmented urothelium. Due to the large class imbalance in the combined WHO'73 and WHO'04 grading schemes, the classification network focuses on the classification according to the WHO'04 system. The data were divided in three sets on a patient level. About 60% of the data were used for training of both networks (training set), 20% for selection of the best performing network (validation set), and 20% was to assess the performance of the networks (test set). A weighted distribution of annotated pixels of the combined WHO'73 and WHO'04 grading scheme was pursued in all three sets.

Segmentation Network

Patches of the urothelium as delineated by the expert observers and containing tumorous urothelium and

Figure 1 A and B: Example of a non-muscle-invasive bladder cancer (A) with the corresponding delineations (B). C: A detailed image of an excluded region caused by cauterization artifacts. D and E: A close-up of the tissue and the corresponding delineation. Original magnification, $\times 20$.

non-atypical urothelium were used as input. The patch size was 572×572 RGB pixels, which corresponds to $286 \times 286 \mu\text{m}$, with an overlap of 25% in the patches. To reduce the computational demand, the background pixels of each image were automatically detected. Only patches with $>25\%$ of foreground pixels were included. The segmentations of the patches resulting from the U-Net were then stitched together to obtain the segmentation of the whole tissue sample.

Weighted cross-entropy was adopted as loss function to accurately find the urothelium. Using this, a high penalty (3.5 times higher than for normal tumor tissue) was given for missing regions with non-atypical urothelium, and no penalty was given for regions that were marked as nondiagnostic by the expert observer. The Adam optimizer²⁰ was used to optimize the performance of the U-Net, with a learning rate of 0.0004 ($\beta_1 = 0.9$, $\beta_2 = 0.999$). A dropout of 50% was used to prevent overfitting. Image augmentation was applied to make the network more robust to variation of the input images. The applied augmentations were random color variation, flipping, and mirroring of the training patches.

Assessment of Accuracy

The evaluation of the segmentation network was based on a qualitative assessment of whether the diagnostic urothelium as annotated by the expert observer was detected. All other regions that were segmented by the segmentation network

were visually checked to determine whether these regions contained urothelium. The false-positive regions and false-negative regions have been reported for the test set.

Classification Network

The urothelium regions as determined by the segmentation network were used for the training of the second network, namely the classification network. This network used the ImageNet pretrained 16-layer VGG architecture and was optimized for the grading of bladder tissue according to the WHO'04 grading system.

Patch Generation

The classification network only processed regions in which urothelium was detected by the segmentation network. Patches of 224×224 RGB pixels (corresponding to $112 \times 112 \mu\text{m}$), with 25% of overlap, were extracted from these regions. These patches were subdivided in three different categories: i) undefined, ii) WHO'04 low grade, and iii) WHO'04 high grade.

All regions that contained urothelium as determined by the segmentation network and were annotated by the expert observers were labeled according to the consensus reading of the three pathologists (C.D.S.-H. and two non-author pathologists). All regions that were recognized by the segmentation network as urothelium but were not delineated by

the expert observers were placed in the undefined class. Although these regions were correctly classified as being urothelium, the pathologists were unable to assess the grade of the tissue due to the artifacts. The misclassification of patches for the undefined category was not penalized. If there were two different categories present in a single patch according to the network, the patch was classified according to the dominant category, meaning that the dominant category was at least five times more present than the other category. If this selection criterion was not met, the patch was excluded from the classification network.

Training of the Convolutional Neural Network

To adapt the ImageNet pretrained 16-layer VGG architecture for our three-category classification problem, the fully connected and dense layers of the 16-layer VGG architecture were trained from scratch in combination with a dropout of 50%. These classification layers were trained for four epochs, while maintaining the rest of the convolutional layers as frozen. After those epochs, the last two convolutional blocks together with the classification layers were optimized for our classification problem for several epochs, while fixing the ImageNet feature maps of the earlier layers. The network was trained on 224×224 RGB pixel patches in the Keras framework version 2.1.6 using TensorFlow version 1.8.0 as backend (Google Brain, Mountain View, CA). The softmax function was used as activation function of the fully connected layer. The optimizer used was Adam,²⁰ with a learning rate of 0.0005 ($\beta_1 = 0.99$, $\beta_2 = 0.999$). The training patches were augmented by randomly flipping and mirroring to increase the robustness of the network. The epoch resulting in a network with the lowest weighted categorical cross-entropy loss was selected as the best performing network, in which no weight was assigned to patches in the undefined category. The output of the network is the probability of each patch belonging to all three classes, and these probabilities were extracted for both the validation set and the test set. The sum of these three categories by definition equaled one.

Assessment of the Accuracy of the Classification Network

Patches with a probability of the undefined category higher than 0.2 were excluded from the analysis. A receiver-operating characteristic curve for the validation set was then generated. The maximal Youden index was selected as the optimal threshold for the grading of the patches. The accuracy was evaluated on the test set. By clustering the patches at case level, the majority of the classified patches defined the grade of the case. This case level was then used to determine the sensitivity, specificity, and accuracy for the complete test set.

A patient could have multiple cases in the study, either caused by the recurrence of the tumor, or because multiple

tumors were excised in one procedure. For evaluation of the performance, slides with bad staining or only small regions with diagnostic urothelium were excluded. The performance was assessed by using a linear weighted κ , accuracy, sensitivity, and specificity. Subsequently, the performance was assessed and compared with the agreement of pathologists with the consensus reading. Moreover, the Fleiss κ was determined to assess the degree of agreement, among the pathologists and pathologists with automated grading. Levels of agreement for κ between 0.21 and 0.40 were assigned as fair agreement, 0.41 to 0.60 as moderate agreement, 0.61 to 0.80 as substantial agreement, and 0.81 to 1.00 as good agreement.

Results

The annotation of the data set resulted in a total of 32 billion annotated pixels, of which only 51 million pixels contained non-atypical urothelium. The number of annotated pixels according to the combined WHO'73 and WHO'04 grading scheme was unequally distributed among the categories (Table 1). This resulted in roughly 0.5 million patches for the segmentation network.

Segmentation Network

The U-Net detected more urothelium than was annotated by the expert observer. Therefore, the accuracy of the segmentation network cannot be expressed in terms of overlap between the delineations and segmentation as a result of the U-Net. The three most common findings of nondiagnostic urothelium were: i) von Brunn's nests, which are solid nests of benign urothelium in the lamina propria, ii) urothelium within regions of mechanical damage, such as artifacts caused by cauterization or crush artifacts, and iii) out-of-focus regions of the section. In 13% of the samples of the test set, false-positive regions were noted. Those regions were mainly found in slides with extensive color loss or in regions of inflammation. Similar performance was observed in the training set and in the validation set.

Examples of the performance of the segmentation network on a poorly stained section and a strongly stained section are illustrated in Figure 2. Urothelium was detected in both cases; however, the border of the urothelium is less well defined in the poorly stained section.

Classification Network

The regions segmented by the segmentation network resulted in a total of 1.2 million patches for the classification network, of which roughly 680,000 were used for training, 220,000 for selecting the best performing network and optimizing the hyperparameters, and 290,000 patches were used to assess the performance of the network. There is a large class imbalance, as only 0.16% of the annotated pixels contain non-atypical tissue, and grade 2/low is almost seven

Table 1 Overview of the Annotated Urothelium for the Segmentation Network

Category	Annotated pixels, <i>n</i>	Annotated pixels, %	Patients, <i>n</i>
Non-atypical	51,829,465	0.16	36
1/low	2,310,733,822	7.14	48
2/low	15,581,924,846	48.14	103
2/high	6,816,413,008	21.06	60
3/high	7,604,248,606	23.50	49
Urothelium	32,365,149,747		232

times more represented than grade 1/low (Table 1). The patches were therefore classified based on the WHO'04 classification and to exclude non-atypical tissue from the classification. An overview of the patches and subdivision over the classes and sets is listed in Table 2.

Assessment of Accuracy of the Classification Network

After training the ImageNet pretrained 16-layer VGG network, the slide quality of the unseen test set was assessed to exclude cases with nonrepresentative staining or cases with very limited regions of diagnostic urothelium. This occurred under supervision of specialized pathologists (C.D.S.-H.), excluding only cases that were clinically non-diagnostic. The performance shows that the automated grading exhibits moderate agreement ($\kappa = 0.48 \pm 0.14$ SEM) with the consensus reading (Table 3). The agreement among pathologists (C.D.S.-H.) ranges between fair ($\kappa = 0.35 \pm 0.13$ SEM and $\kappa = 0.38 \pm 0.11$ SEM) and moderate ($\kappa = 0.52 \pm 0.13$ SEM). The Fleiss κ coefficient among the three pathologists (C.D.S.-H.) was 0.39 ± 0.09 SEM, whereas the reliability was nonsignificantly higher for

the automated grading (0.40 ± 0.06 SEM). The automated classification correctly graded 76% of the low-grade cancers according to the consensus reading and 71% of the high-grade cancers.

Discussion

This study presents a methodology for automatic segmentation and grading of urothelial tumors by the use of deep learning. This approach could be the first step toward reliable and reproducible grading of bladder tumors. We trained and evaluated a first network focused on the segmentation of urothelium. The output of this network was then used to train the grading network. The automated grading was then compared with the grading of three experienced pathologists.

Two earlier studies have used machine learning techniques for the automated grading of bladder cancer, using histologic and nuclear features from specific regions of interest.^{13,14} Herein an automated method to grade bladder cancer is presented that differs from earlier studies by reducing the need of human input and thereby increasing the reproducibility.^{13,14} Furthermore, the performance of the automated grading is in line with other recently published studies. Soukup et al⁶ reported an agreement ranging from 65% to 88% with κ values between 0.30 and 0.73 for the classification of low- and high-grade tumors following the WHO'04 guidelines. Another study reported an interobserver variability of 27% to 63% for low-grade tumors and 21% to 67% for high-grade tumors.²¹ Although a moderate agreement may be considered insufficient for the implementation in clinical practice, it should be noted that the findings of other interobserver studies are in agreement with the performance of our automated grading methodology.^{6,22–24}

Figure 2 A, B, F, and G: Example of the performance of the segmentation network on a poorly stained section (A and F) and on a well-stained section (B and G). C and H: The detection of urothelium with mechanical damage is shown. D and I: The correct segmentation of non-atypical urothelium is shown. E and J: An example of a false-positive detected region can be seen, in which both mechanical damage and inflammation are present. The dark gray regions in F–J are automatically detected as urothelium by the segmentation network. Scale bars = 100 μ m (A–J).

Table 2 Extracted Patches of Automated Segmented Urothelium for the Classification Network

Category	Patches, <i>n</i> (train–validation–test)
Undefined	63,308–27,112–32,712
WHO’04 Low grade	310,562–121,160–132,988
WHO’04 High grade	302,120–71,556–119,698
Total	675,990–219,828–285,398

WHO, World Health Organization.

The current study has several strengths. First, this study contained TURBT specimens originating from three different hospitals; second, the digitized slides that were included contained both old and recent sections. All cases were assessed by three experienced pathologists with a special interest in uropathology, and a consensus diagnosis was formulated. Furthermore, the difference in staining caused by the institutional staining protocol^{25–27} and by the fading of the older sections resulted in a first step toward a generic model.

The current classification method relies on two individual neural network architectures. The first network, segmentation, focuses on the detection of urothelium. The second network, aimed at classification, uses this input for grading the digitized sections. However, no direct feedback from the second network was incorporated into the first network. If this connection is made, the performance of both networks could be optimized at the same time.^{28–30} The single-way communication between the segmentation and classification networks facilitates the use of the segmentation network as a simple region-of-interest detector for bladder cancer images.

The classification network assessed the tumor grade on a case level, based on the majority vote of the classification of patches, assuming bladder cancer to be a homogeneous tumor. Borders of low- and high-grade regions are not clearly demarcated, and making differentiation on a patch level is therefore likely to increase the interobserver variation even more. This case-level grading flattens out the effect of potential patches that contain areas with out-of-focus tissue, inflammation, or mechanical artifacts. However, clinical guidelines advise grading a heterogeneous

tumor on the highest grade observed,¹¹ and the assessment of the pathologists was based on the highest grade pathologic pattern present on a patient level. However, compared with glandular tumors such as prostate cancer, UCC is a more homogeneous disease. In future research, approaches accounting for heterogeneity in tumors (eg, as suggested for the automated Gleason grading of prostate biopsy specimens) will be considered,^{16,31} with both studies relying on case-level classification only. Both approaches as proposed by Bulten et al³¹ and Campanella et al¹⁶ require significantly more slides than were available for this study. When similar amounts of data become available for NMIBC, further differentiation can be made between the grading; for example, by incorporating the WHO’73 grading as well, which still has a substantial role in clinical decision-making.^{2,12} The WHO’16 grading system was recently introduced, to be used in extension of the WHO’04 system. However, no substantial differences exist between the WHO’04 and the WHO’16 systems.^{5,6}

For the grading of UCC, a magnification of 10× to 20× is advised according to the guidelines.¹¹ Therefore, in the current study, the digitized slides were exported at a magnification of 20×. However, in future studies, a magnification of 40× may be considered, allowing better insight into smaller features.

Because the data set from the current study consisted of multi-institutional data, and relatively old sections were included, staining differences were prominent. Several studies have been conducted regarding color normalization,³² and it has been shown that color normalization increases diagnostic performance on data sets from different institutions.³³ However, the deep learning approach was trained on a very diverse data set, making the network more robust for those color fluctuations. Small improvements might be gained by color normalization, but this approach would severely increase the computational complexity.

Overfitting is a major risk when training a deep learning architecture on a small data set. Although the current data set consisted of >1 million image patches for the automated grading, these patches were originating from only 232 patients. Although this number is considered a small sample

Table 3 Performance Measures of the Agreement between Pathologists, the Consensus Reading, and the Automated Grading

Comparison	κ , means $\pm$ SEM	Accuracy, %	Sensitivity, %	Specificity, %
Automated vs consensus	0.48 $\pm$ 0.14	74	71	76
Observer 1 vs consensus	0.38 $\pm$ 0.12	69	100	38
Observer 2 vs consensus	0.81 $\pm$ 0.09	91	91	91
Observer 3 vs consensus	0.62 $\pm$ 0.12	81	86	76
Automated vs observer 1	0.35 $\pm$ 0.11	67	100	36
Automated vs observer 2	0.38 $\pm$ 0.14	69	70	68
Automated vs observer 3	0.48 $\pm$ 0.13	74	80	68
Observer 1 vs observer 2	0.38 $\pm$ 0.11	69	63	100
Observer 1 vs observer 3	0.35 $\pm$ 0.13	69	65	88
Observer 2 vs observer 3	0.52 $\pm$ 0.13	76	81	71

size, this study is the largest to date regarding automated grading of UCC using deep learning. To reduce the risk of overfitting, an ImageNet pretrained architecture was used. Of this network, only the two last convolutional blocks were fine-tuned for the histopathology data. This approach reduces the number of trainable features and therefore prevents the network from overfitting. Together with the use of a 50% dropout rate during the training stage, the risk of overfitting was most likely reduced. However, overfitting cannot be completely ruled out without the use of a large external validation set.

Histopathologic grading by pathologists is subjective to interobserver variability. To reduce this interobserver variation, an automated segmentation and grading method for NMIBC was proposed and implemented. Automated methods, such as described in this study, are not susceptible to subjectivity or fatigue, and offer a route to a reliable and consistent grading procedure. Further efforts could be made to increase the reliability of the methodology, starting by accumulating more data from different institutions.

Because the performance of the deep learning network can only be as good as the gold standard, it is impossible to perform better than the consensus of the pathologists. Therefore, disease outcome could be a better end point when training a deep learning network. However, this introduces numerous confounders, such as treatment regimen and baseline characteristics.

In conclusion, this study showed that it is possible to automatically detect and grade NMIBC with an accuracy comparable to that of pathologists by combining a U-Net segmentation and classification network. The segmentation network was used as a region-of-interest detector, whereas the segmentation network provided an objective opinion for agile clinical decision-making.

Acknowledgment

We thank NVIDIA Corporation for the donation of a Titan X GPU card for our research.

References

- Babjuk M, Böhle A, Burger M, Capoun O, Cohen D, Compérat EM, Hernández V, Kaasinen E, Palou J, Rouprêt M, van Rhijn BWG, Shariat SF, Soukup V, Sylvester RJ, Zigeuner R: EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: update 2016. *Eur Urol* 2017, 71:447–461
- Sylvester RJ, van der Meijden APM, Oosterlinck W, Witjes JA, Bouffoux C, Denis L, Newling DWW, Kurth K: Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol* 2006, 49:466–477
- Miyamoto H, Miller JS, Fajardo Da, Lee TK, Netto GJ, Epstein JI: Non-invasive papillary urothelial neoplasms: the 2004 WHO/ISUP classification system. *Pathol Int* 2010, 60:1–8
- Mostofi FK, Davis CJ, Sesterhenn IA, Sobin LH: *Histological Typing of Urinary Bladder Tumours*. Geneva, Switzerland, World Health Organization, 1999
- Humphrey PA, Moch H, Cubilla AL, Ulbright TM, Reuter VE: The 2016 WHO classification of tumours of the urinary system and male genital organs—part B: prostate and bladder tumours. *Eur Urol* 2016, 70:106–119
- Soukup V, Čapoun O, Cohen D, Hernández V, Babjuk M, Burger M, Compérat E, Gontero P, Lam T, MacLennan S, Mostafid AH, Palou J, van Rhijn BWG, Rouprêt M, Shariat SF, Sylvester R, Yuan Y, Zigeuner R: Prognostic performance and reproducibility of the 1973 and 2004/2016 World Health Organization grading classification systems in non-muscle-invasive bladder cancer: a European Association of Urology Non-muscle Invasive Bladder Cancer Guidelines Panel systematic review. *Eur Urol* 2017, 72:801–813
- Malmström PU, Busch C, Johan Norlén B: Recurrence, progression and survival in bladder cancer. *Scand J Urol Nephrol* 1987, 21:185–195
- Pauwels RP, Schapers RF, Smeets AW, Debruyne FM, Geraedts JP: Grading in superficial bladder cancer. (1). Morphological criteria. *Br J Urol* 1988, 61:129–134
- Epstein JI: The new World Health Organization/International Society of Urological Pathology (WHO/ISUP) classification for TA, T1 bladder tumors: is it an improvement? *Crit Rev Oncol Hematol* 2003, 47:83–89
- World Health Organization: *World Health Organization Classification of Tumours*. Edited by Eble JN, Sauter G, Epstein JI, Sesterhenn IA. In *Pathology and Genetics of Tumours of the Urinary System and Male Genital Organs*. Lyon, France: IARC Press, 2004. pp. 217–278
- World Health Organization. *WHO Classification of Tumours of the Urinary System and Male Genital Organs*. Edited by Moch H, Humphrey PA, Ulbright TM, Reuter VE. Lyon, France: International Agency for Research on Cancer, 2016
- Fernandez-Gomez J, Madero R, Solsona E, Unda M, Martinez-Piñeiro L, Gonzalez M, Portillo J, Ojea A, Pertusa C, Rodriguez-Molina J, Camacho JE, Rabadan M, Astobieta A, Montesinos M, Isorna S, Muntanola P, Gimeno A, Blas M, Martinez-Piñeiro JA: Predicting nonmuscle invasive bladder cancer recurrence and progression in patients treated with bacillus Calmette-Guerin: the CUETO scoring model. *J Urol* 2009, 182:2195–2203
- Spyridonos P, Cavouras D, Ravazoula P, Nikiforidis G: Neural network-based segmentation and classification system for automated grading of histologic sections of bladder carcinoma. *Anal Quant Cytol Histol* 2002, 24:317–324
- Choi H, Jarkrans T, Bengtsson E, Vasko J, Wester K, Malmström PU, Busch C: Image analysis based grading of bladder carcinoma. Comparison of object, texture and graph based methods and their reproducibility. *Anal Cell Pathol* 1997, 15:1–18
- Litjens G, Sánchez CI, Timofeeva N, Hermesen M, Nagtegaal I, Kovacs I, Hulsbergen-van de Kaa C, Bult P, van Ginneken B, van der Laak J: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 2016, 6:1–11
- Campanella G, Silva VWK, Fuchs TJ: Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arXiv*, 2018. [EPub] <https://arxiv.org/abs/1805.06983>
- Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM; CAMELYON16 Consortium: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017, 318:2199–2210
- Boschieter J, Hentschel A, Savci-Heijink CD, van der Voorn JP, Rozendaal L, Vis AN, van Rhijn BWG, Lissenberg-Witte BI, van de Putte EEF, van Moorselaar RJA, Nieuwenhuijzen JA: Reproducibility and prognostic performance of the 1973 and 2004 World Health Organization classifications for grade in non-muscle-invasive bladder cancer: a multicenter study in 328 bladder tumors. *Clin Genitourin Cancer* 2018, 16:e985–e992
- Kamphuis G, de Bruin D, Brandt M, Knoll T, Conort P, Lapini A, Dominguez-Escrig J, de la Rosette JJMCH: Comparing image perception of bladder tumours in four different storz professional

- image enhancement system (SPIES) modalities using the iSPIES app. *J Endourol* 2016, 30:1–20
20. Kingma DP, Ba JL: Adam: a method for stochastic optimization. arXiv, 2017. [EPub] <https://arxiv.org/abs/1412.6980>
 21. Comprat EM, Burger M, Gontero P, Mostafid AH, Palou J, Rouprt M, van Rhijn BWG, Shariat SF, Sylvester RJ, Zigeuner R, Babjuk M: Grading of urothelial carcinoma and the new “World Health Organisation Classification of Tumours of the Urinary System and Male Genital Organs 2016.” *Eur Urol Focus* 2019, 5:457–466
 22. May M, Brookman-Amissah S, Roigas J, Hartmann A, Strkel S, Kristiansen G, Gilfrich C, Borchardt R, Hoshke B, Kaufmann O, Gunia S: Prognostic accuracy of individual uropathologists in noninvasive urinary bladder carcinoma: a multicentre study comparing the 1973 and 2004 World Health Organisation classifications. *Eur Urol* 2010, 57:850–858
 23. van Rhijn BWG, van Leenders GJLH, Ooms BCM, Kirkels WJ, Zlotta AR, Boev ER, Jbsis AC, van der Kwast TH: The pathologist’s mean grade is constant and individualizes the prognostic value of bladder cancer grading. *Eur Urol* 2010, 57:1052–1057
 24. Babjuk M, Burger M, Comprat E, Gontero P, Mostafid AH, Palou J, van Rhijn BWG, Rouprt M, Shariat SF, Sylvester R, Zigeuner R: EAU guidelines on non-muscle-invasive bladder cancer (TaT1 and CIS). *Eur Assoc Urol* 2018:11–12, 17–19
 25. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, Schmitt C, Thomas NE: A method for normalizing histology slides for quantitative analysis Proceedings of the International Symposium on Biomedical Imaging (ISBI). From Nano Macro, ISBI; 2009. pp. 1107–1110
 26. Krishnamurthy S, Mathews K, McClure S, Murray M, Gilcrease M, Albarracin C, Spinosa J, Chang B, Ho J, Holt J, Cohen A, Giri D, Garg K, Bassett RL, Liang K: Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin-stained breast tissue sections. *Arch Pathol Lab Med* 2013, 137:1733–1739
 27. Onder D, Zengin S, Sarioglu S: A review on color normalization and color deconvolution methods in histopathology. *Appl Immunohistochem Mol Morphol* 2014, 22:713–719
 28. Mehta S, Mercan E, Bartlett J, Weaver D, Joann G, Shapiro L: Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. arXiv, 2018. [EPub] <https://arxiv.org/abs/1806.01313>
 29. Girard F, Kavalec C, Cheriet F: Joint segmentation and classification of retinal arteries/veins from fundus images. *Artif Intell Med* 2019, 94: 96–109
 30. Shen F, Gan R: Joint segmentation and classification with fully convolutional networks. 2016 3rd Int Conf Syst Informatics. ICSAI 2016. pp. 338–343
 31. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, van der Laak J, de Kaa CH, Litjens G: Automated Gleason grading of prostate biopsies using deep learning. arXiv, 2019. [EPub] <https://arxiv.org/abs/1907.07980>
 32. Anghel A, Stanisavljevic M, Andani S, Papandreou N, Rschoff JH, Wild P, Gabrani M, Pozidis H: A high-performance system for robust stain normalization of whole-slide images in histopathology. *Front Med* 2019, 6:1–13
 33. Bejnordi BE, Litjens G, Timofeeva N, Otte-Hller I, Homeyer A, Karssemeijer N, van der Laak JAWM: Stain specific standardization of whole-slide histopathological images. *IEEE Trans Med Imaging* 2016, 35:404–415