THE LANCET
Oncology

Supplementary appendix

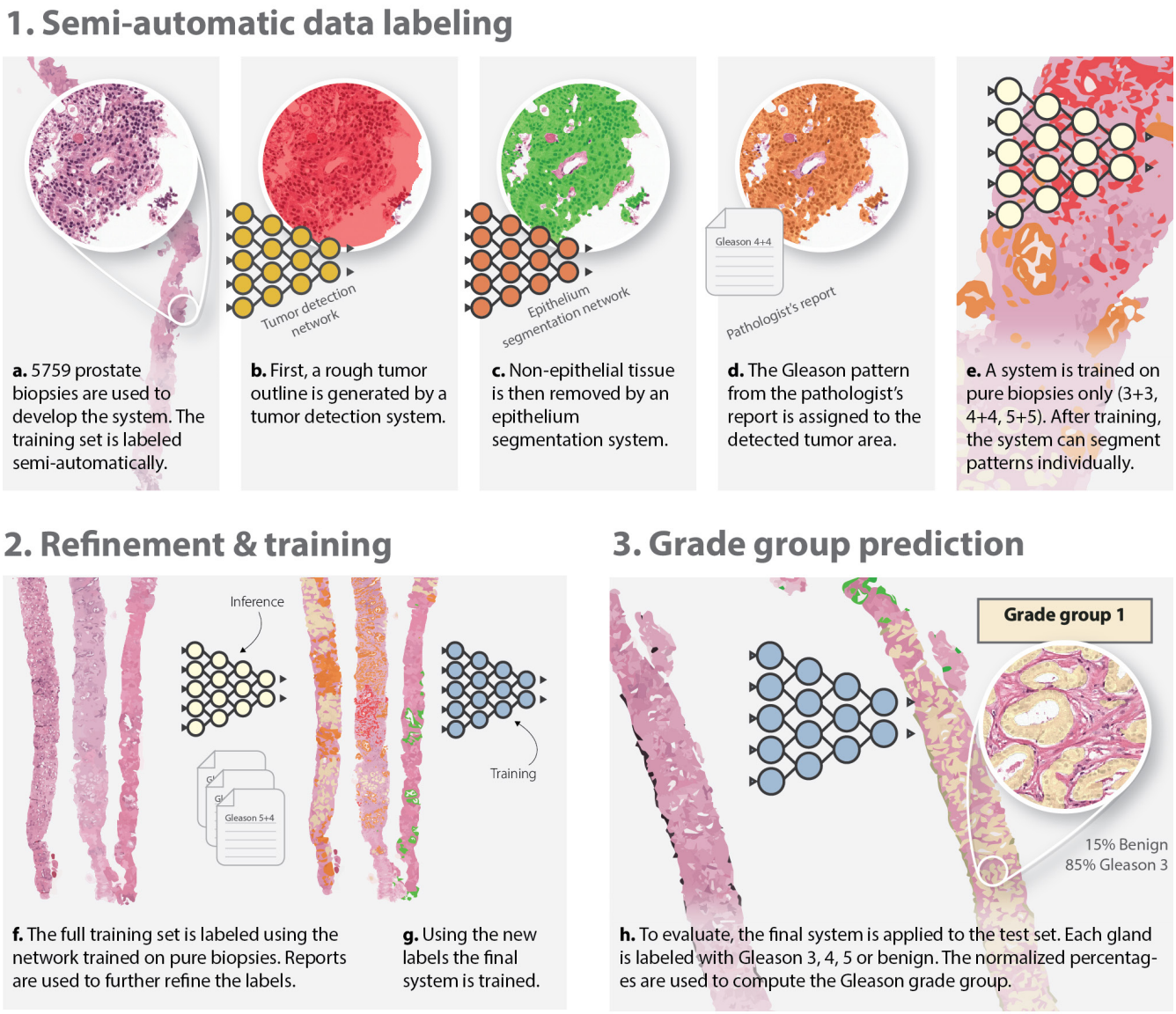
This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 2020; published online Jan 8. https://doi.org/10.1016/S1470-2045(19)30739-9.

Contents

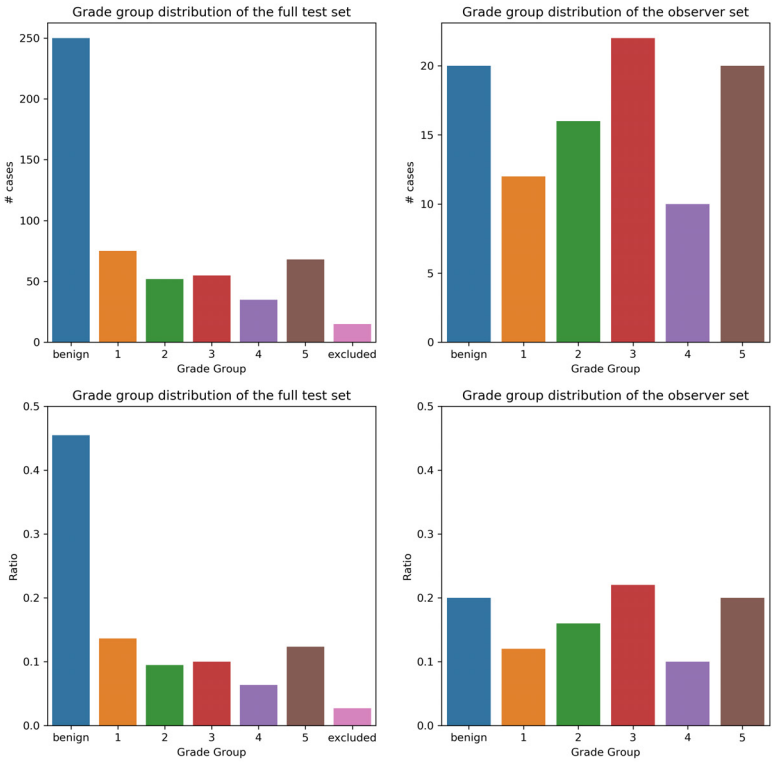
1 Supplementary figures

1.1 Visual overview of the deep learning method



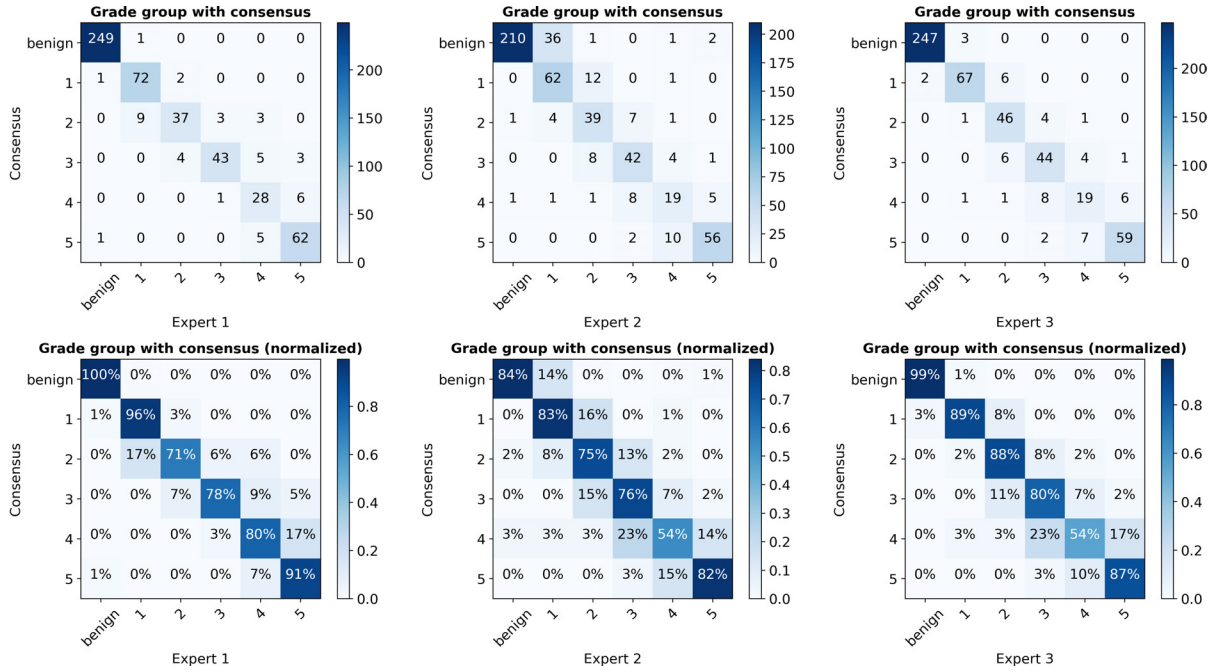
Supplementary Figure 1. Overview of the development of the deep learning system. We employ a semi-automated method of labelling the training data (a-e, top row), removing the need for manual annotations by pathologists. The final system can assign Gleason growth patterns on a cell-level.

1.2 Case distribution test set



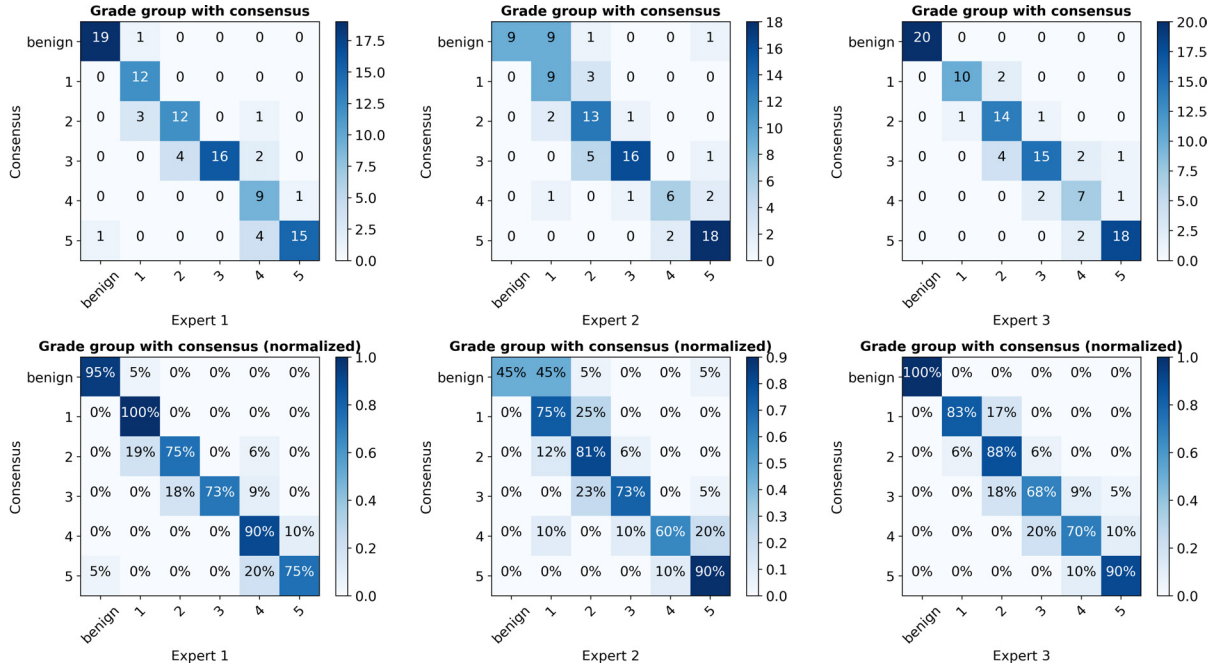
Supplementary Figure 2. Grade group distribution of the cases of the full test set (left) and the observer set that was presented to the external panel (right). Cases for the observer set were sampled from the test set. Top row shows the number of cases in each set, the bottom row the ratio. Both sets differ in distribution: for the full test we opted for a distribution that is close to clinical practice by including a large set of negative cases. For the observer set we included an equal number of cases from all grade groups.

1.3 Confusion matrices experts with consensus on test set



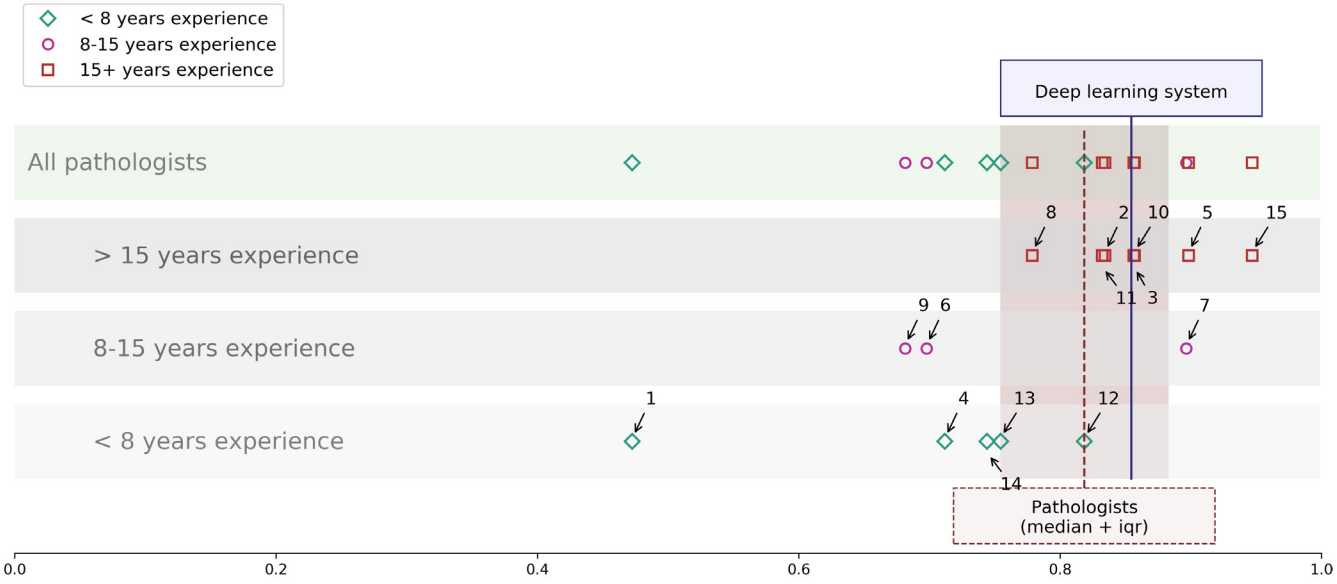
Supplementary Figure 3. Confusion matrices for the three expert pathologists on the test set. The original scores of the pathologists (from round 1) are compared with the final consensus scores.

1.4 Confusion matrices experts with consensus on observer set



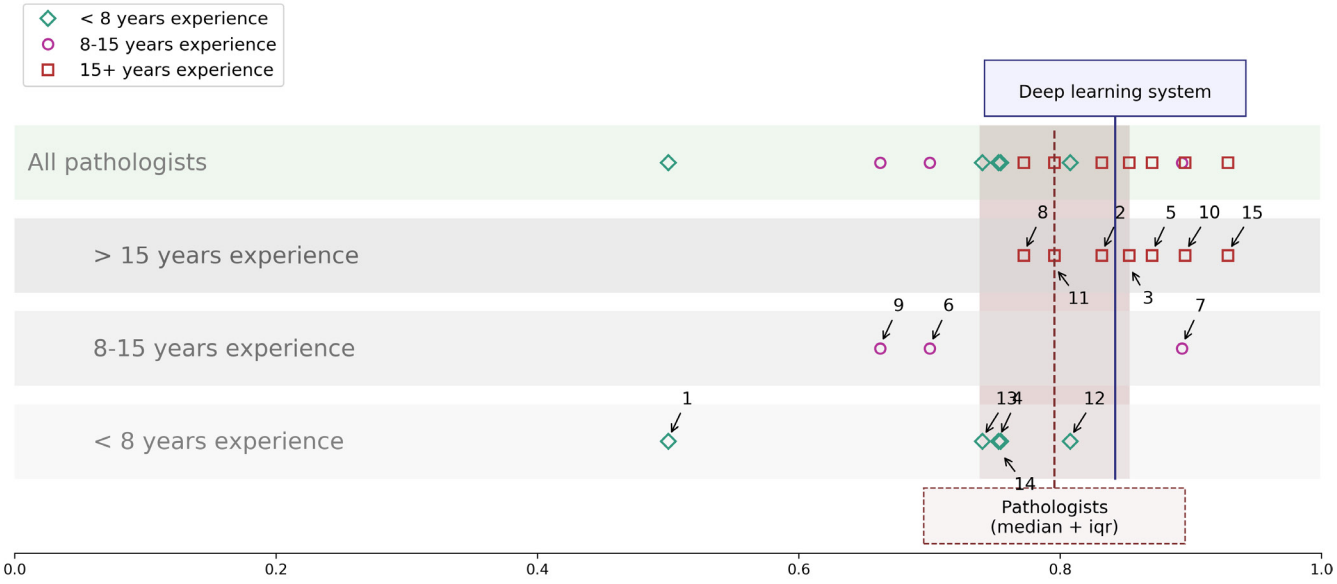
Supplementary Figure 4. Confusion matrices for the three expert pathologists on the observer set. The original scores of the pathologists (from round 1) are compared with the final consensus scores.

1.5 Gleason grade group agreement of deep learning system versus panel



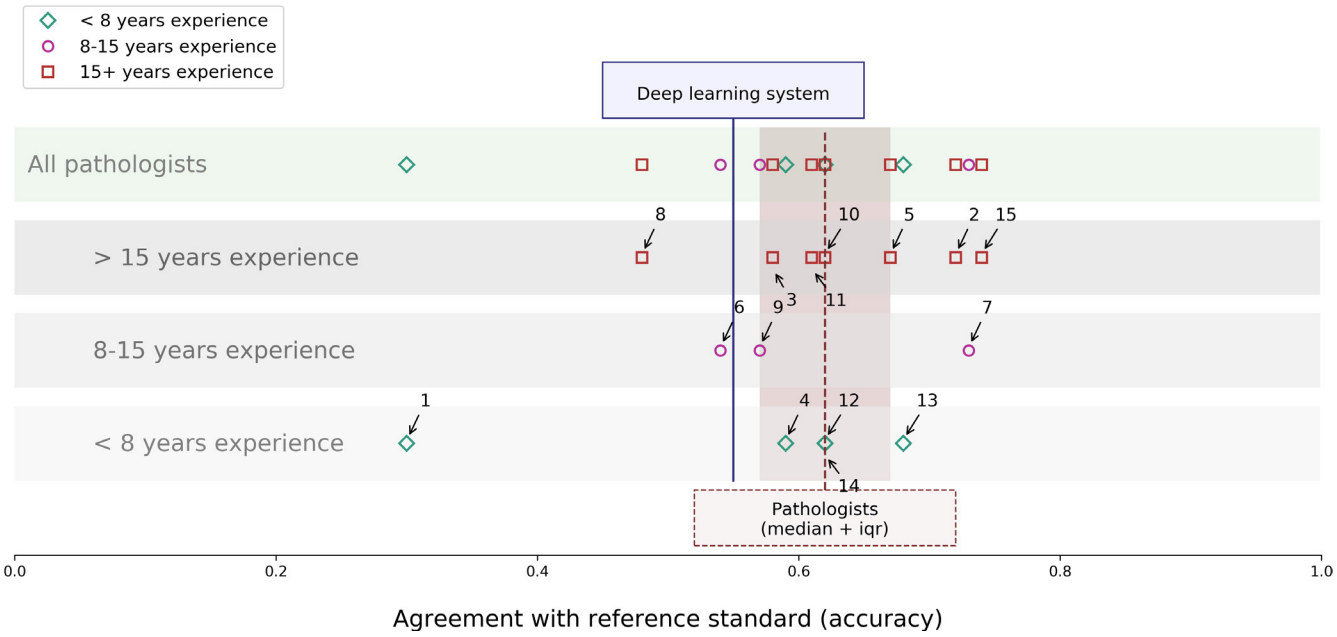
Supplementary Figure 5. Agreement on Gleason grade group for each pathologist of the panel and the deep learning system with the consensus. The panel members are split out according to their experience level. Additionally, the median kappa of the pathologists is shown in brown.

1.6 Gleason score agreement of deep learning system versus panel



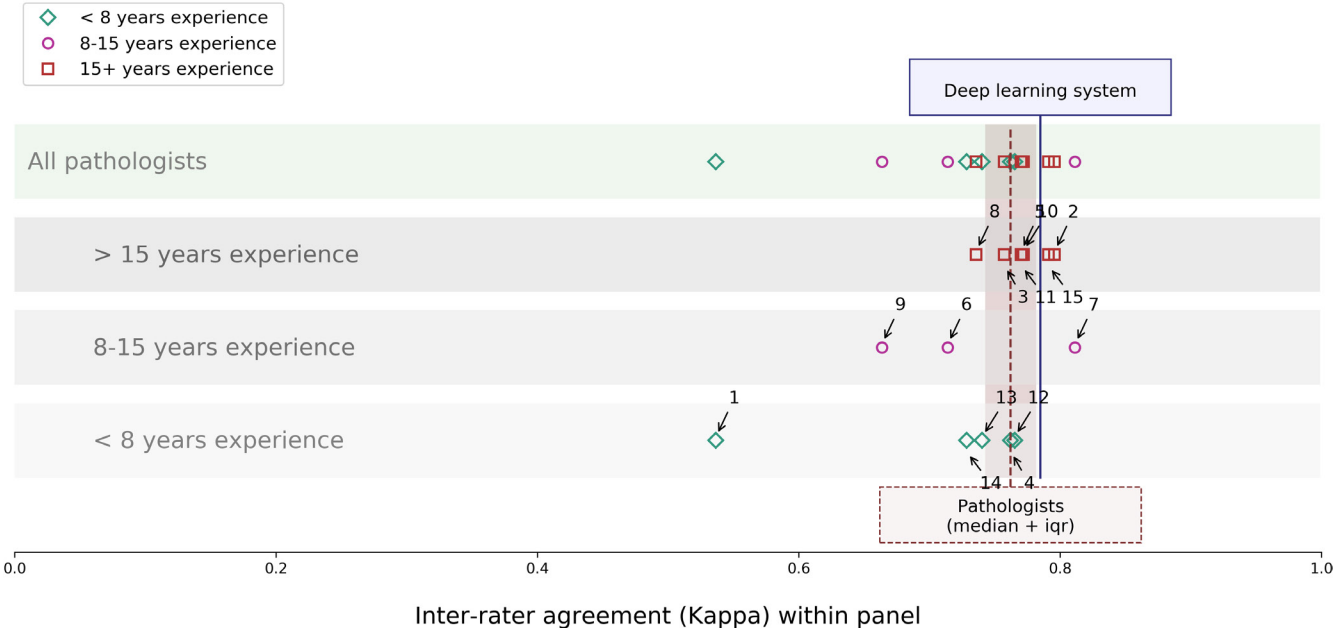
Supplementary Figure 6. Agreement with the reference standard on Gleason score (sum of both primary and secondary pattern) of both the panel members and the deep learning system.

1.7 Accuracy deep learning system accuracy versus panel



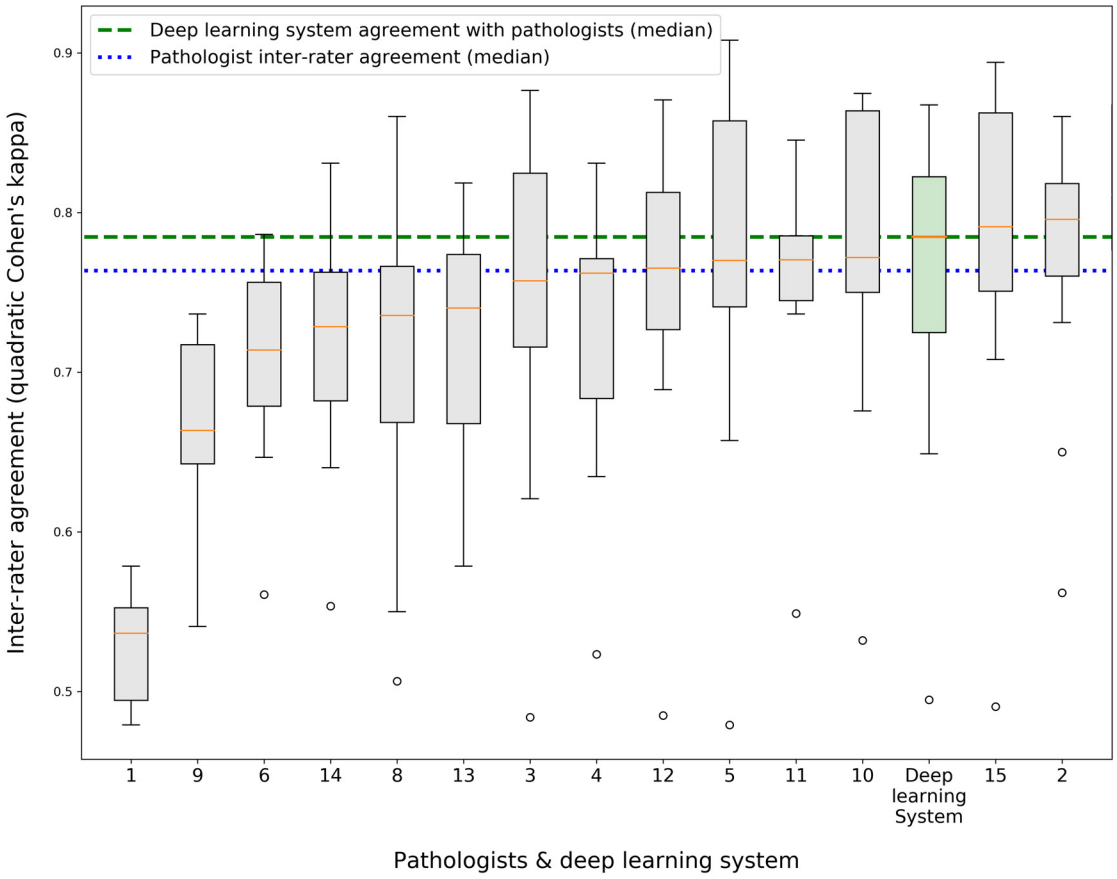
Supplementary Figure 7. Grade group accuracy compared to the reference standard of both the panel members and the deep learning system.

1.8 Grade group agreement of deep learning system versus panel (without reference)



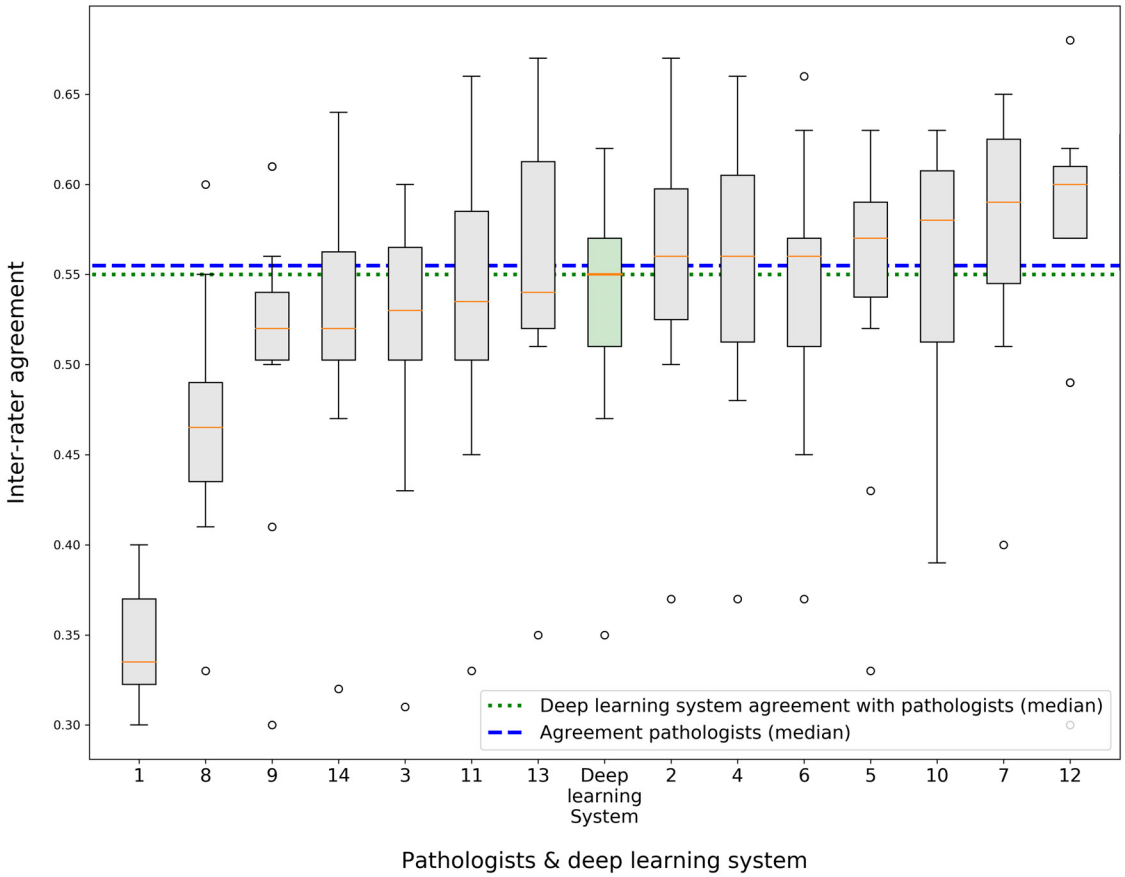
Supplementary Figure 8. Median inter-rater agreement (quadratic kappa) of panel members with each other, compared to the median agreement of the network with the panel. The reference standard, set by the three experts, was not used in this analysis.

1.9 Grade group agreement (quadratic kappa) between pathologists



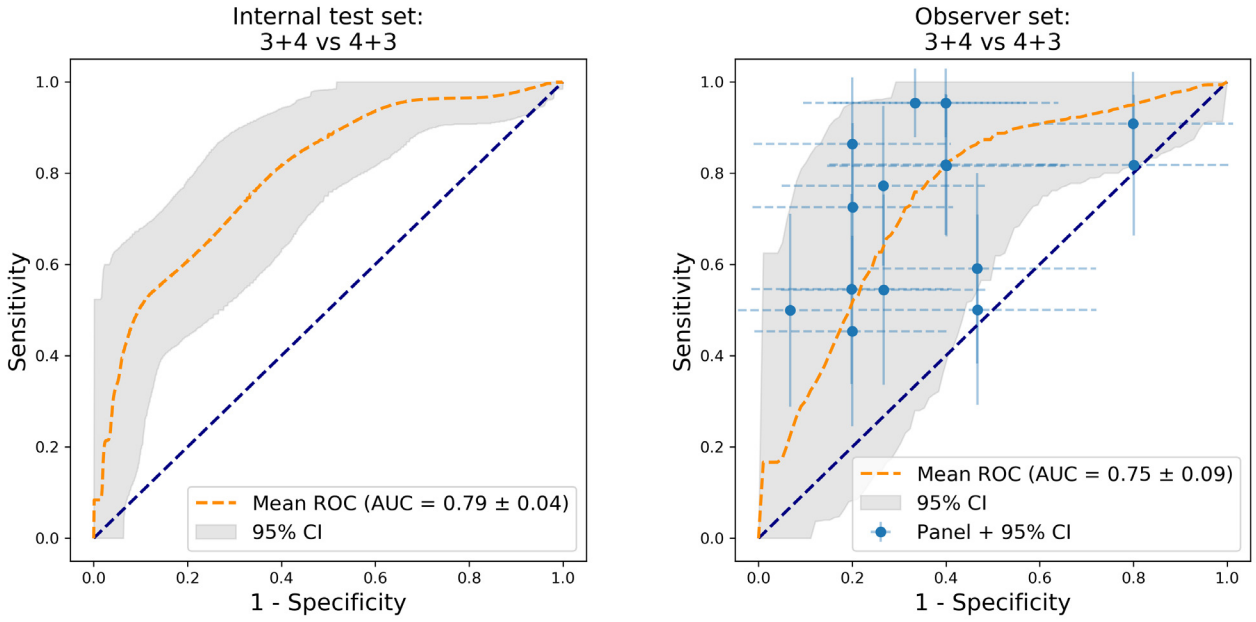
Supplementary Figure 9. Inter-rater agreement between panel members. For each pathologist, the inter-rater agreement with each other pathologist from the panel was calculated. Additionally displayed is the agreement of the deep learning system with the pathologists from the panel. The pathologists and deep learning system are ordered based on their respective median agreement values. The two horizontal lines display the median agreement of all pathologists (in blue) and median agreement of the system (in green).

1.10 Grade group agreement (non-weighted accuracy) between pathologists



Supplementary Figure 10. Inter-rater agreement (non-weighted accuracy measure) between external pathologists. For each pathologist the agreement with each other pathologist was calculated. Additionally displayed is the agreement of the deep learning system with all pathologists.

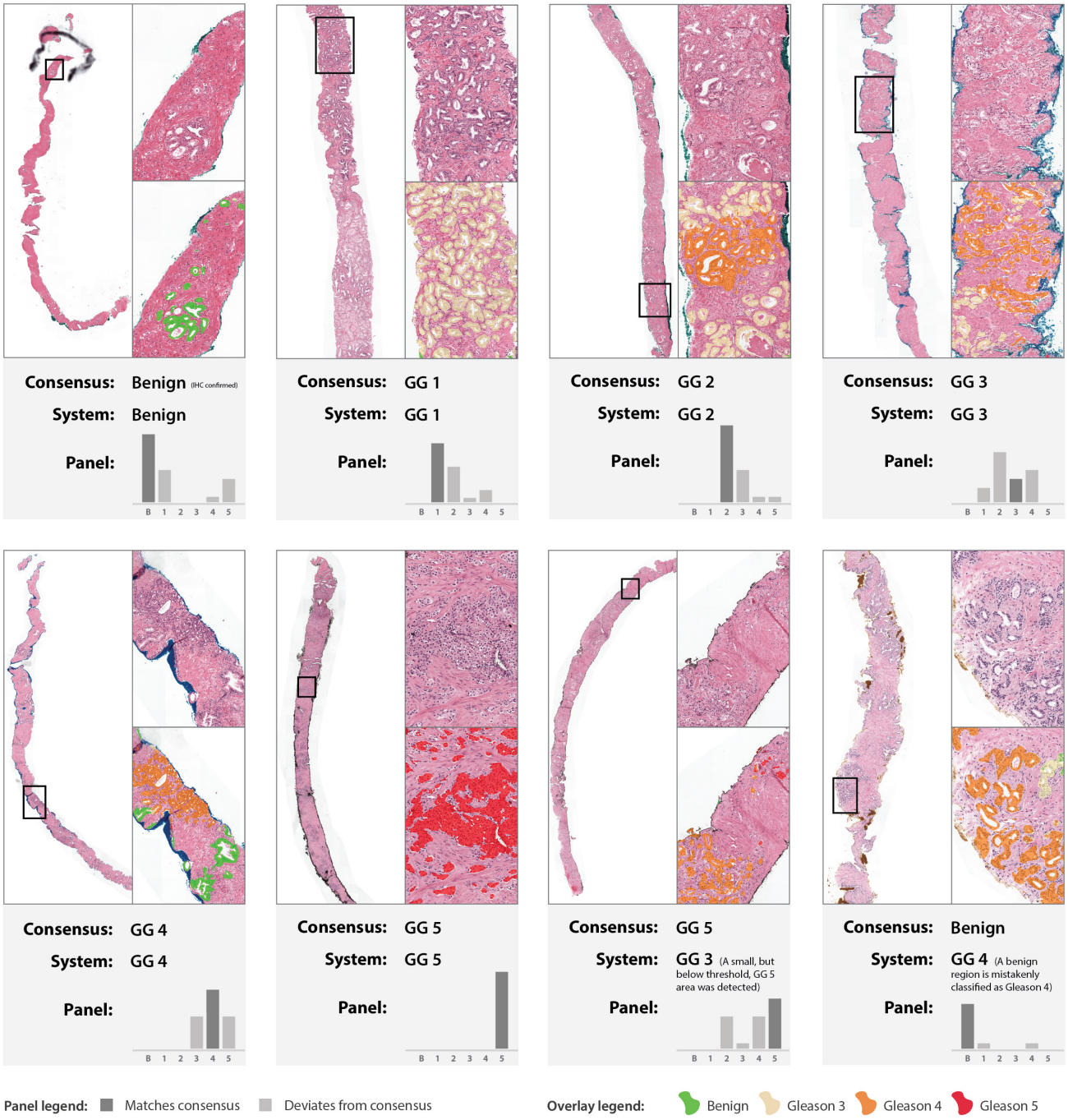
1.11 ROC analysis on Gleason score 3+4 versus 4+3



(a) (b)

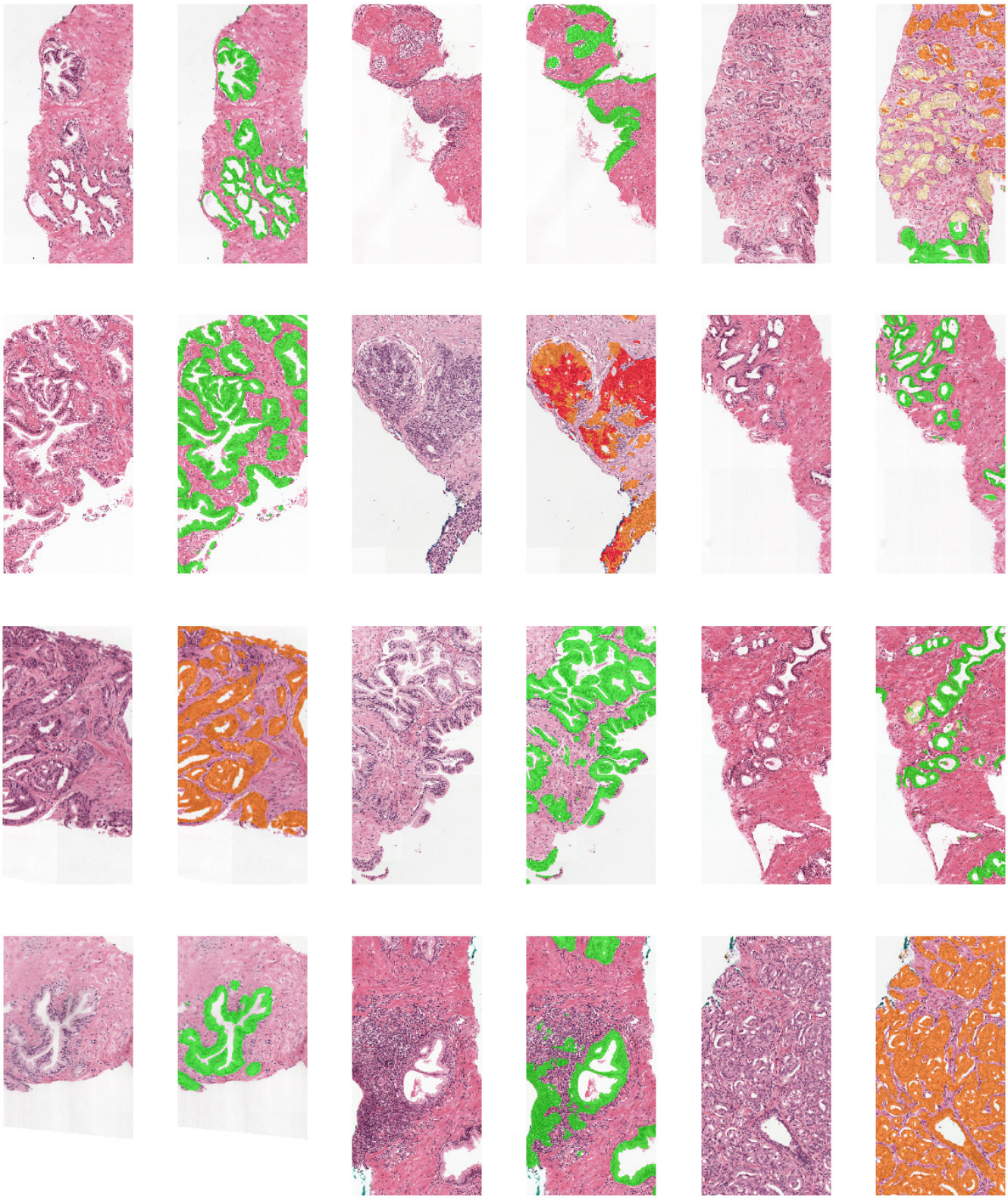
Supplementary Figure 11. ROC analysis on cases with a Gleason score of 3+4 (grade group 2) or 4+3 (grade group 3) for both the internal test set (left) and observer set (right). The deep learning system does not output a continuous prediction value for 3+4 versus 4+3, which is required for a ROC analysis. To still perform the analysis, we used the predicted volume percentage of pattern 4 and 5 as the input of the analysis. To make a fair comparison between the deep learning system and the panel members, panel member Gleason scores higher than 7 were mapped to 4+3, a score of 6 was mapped to 3+4.

1.12 Example cases from observer set, system versus panel



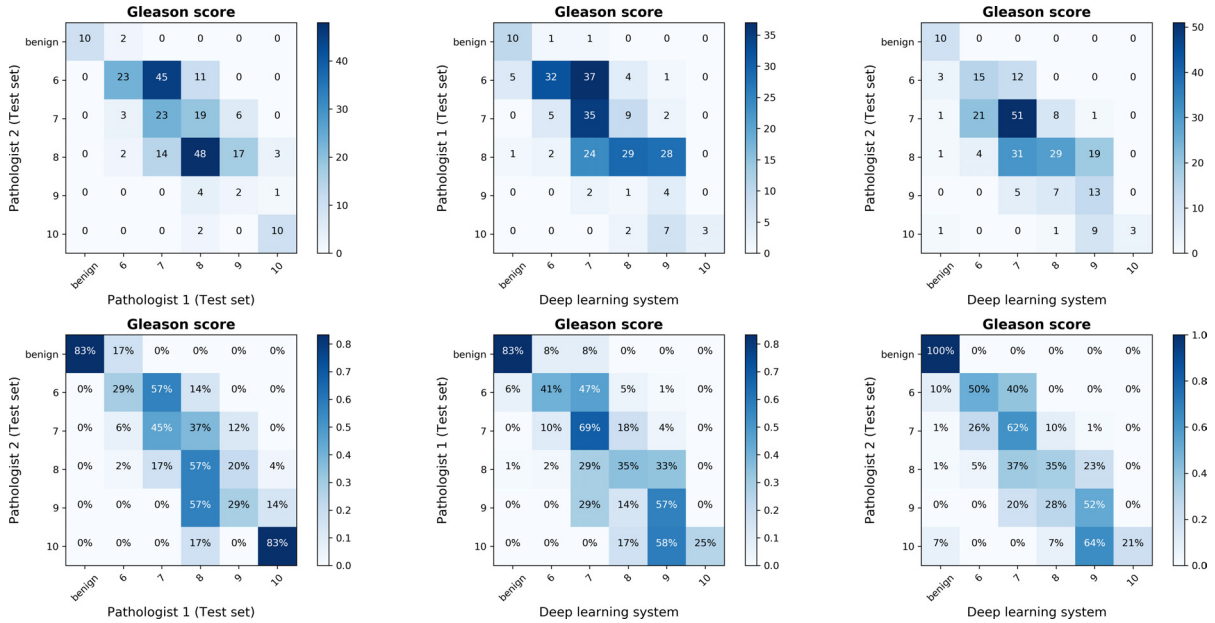
Supplementary Figure 12. Examples from the observer set. For each case, the grade group of the reference standard, the predicted grade group by the deep learning system, and the distribution of grade groups from the panel are shown. In the zoomed regions the Gleason prediction of the system is shown as an overlay on the tissue. The system labels each epithelial gland with either benign (green), Gleason 3 (yellow), Gleason 4 (orange) or Gleason 5 (red). The first six cases show examples of each grade group. The last two cases (bottom row) show examples of failure cases where the system predicted the wrong grade group for the case.

1.13 Test set cases with Gleason overlays



Supplementary Figure 13. Randomly selected cases from the test set. The overlay shows the predictions of the deep learning system: benign tissue (green), Gleason 3 (yellow), Gleason 4 (Orange) and Gleason 5 (red).

1.14 Confusion matrix TMA set



Supplementary Figure 14. Confusion matrices for the external test set of tissue micro arrays. The first column shows the agreement between the two pathologists that set the reference. The second and third column show the results of the deep learning system to the first and second pathologist respectively.

2 Supplementary tables

2.1 Dataset overview

Supplementary Table 1. Overview of the datasets used in the development and validation of the deep learning system.

Source	Dataset name	Size	Reference standard	Purpose
Internal dataset	Training set	4712 biopsies	Semi-automatic method	Development of the deep learning system.
	Tuning set	497 biopsies	Semi-automatic method	Validation of the deep learning system during development.
	Test set	550 biopsies	Three expert pathologists in consensus	Validation of the deep learning system after development. Not used during development.
	Observer set	100 biopsies	Three expert pathologists in consensus	Subset of the internal test set. Used to compare the performance of the deep learning system with a panel of pathologists.
External dataset[1]	Training set	641 TMAs	Single pathologist	To tune the decision thresholds of the deep learning system. Not used for development.
	Test set	245 TMAs	Two pathologists	Validation of the deep learning system on external data. Not used for development.

2.2 Excluded cases from test set

Supplementary Table 2. Excluded cases from the test set. 15 cases were excluded from the test set by at least one of the expert pathologists.

Case ID	Reason for exclusion
44	Tumour area to small to grade.
55	Unsharp.
129	Error in loading file.
161	Possibly pretreated, IHC needed.
184	IHC needed.
185	Mechanically damaged.
248	Too small to grade.
247	IHC needed.
250	Need serial sectioning.
281	Image quality too low for grading.
315	Out of focus.
380	Unsharp.
406	Unsharp.
473	Unsharp.
497	Difficult to grade because of small amount of tumor.

2.3 Consensus meeting cases and final consensus score

Supplementary Table 3. Cases from the test set of which there was no consensus after the second round. These cases were discussed in a consensus meeting with the three experts. For some of the cases the experts indicated that IHC would have been ordered in clinical practice.

Case ID	Expert 1	Expert 2	Expert 3	Final consensus label	Comments
14	negative	5+4	5+4	5+4	IHC required to confirm.
21	negative	3+3	negative	negative	
249	negative	5+5	negative	negative	IHC required to confirm.
257	3+5	3+4	3+4	4+3	
333	negative	3+3	negative	negative	
388	3+5	3+4	3+4	3+4	Possible 3+5.
482	3+3	3+3	negative	negative	IHC required to confirm.
2	4+5	4+3	4+4	4+4	
15	4+5	4+5	4+3	4+5	
106	4+5	4+3	4+3	4+3	
138	4+5	4+3	4+3	4+3	
170	negative	5+5	negative	negative	
211	4+5	4+3	4+3	4+3	
227	negative	3+3	negative	negative	
272	negative	3+3	3+3	negative	IHC required to confirm.
284	negative	3+3	negative	negative	
290	3+5	3+3	4+3	3+4	Volume percentage too low to be certain.
306	negative	3+3	3+3	negative	IHC required to confirm.
317	4+5	4+3	4+3	4+3	IHC required to rule out mechanical damage.
356	3+3	3+4	4+3	3+3	Volume percentage too low to be certain.
369	3+5	3+4	3+5	3+4	Possible 3+5.
372	negative	3+3	negative	negative	IHC required to confirm.
386	negative	3+3	3+3	negative	IHC required to confirm.
390	negative	3+3	negative	negative	IHC required to confirm.
414	5+4	4+3	4+3	4+3	
544	4+5	4+3	4+3	4+4	
545	5+4	4+3	4+3	4+5	

2.4 Classification performance metrics ROC analysis (optimized)

Supplementary Table 4. Classification performance metrics for the three datasets based on the best performing threshold of the ROC analysis.

Decision problem	# cases	F1	Accuracy	Precision	Recall	Specificity	NPV
Internal test set							
Benign vs malignant	250 / 285	0.962	0.961	0.971	0.954	0.968	0.949
Benign & grade group 1 vs grade group ≥ 2	325 / 210	0.918	0.935	0.907	0.929	0.938	0.953
Benign & grade group 1-2 vs grade group ≥ 3	377 / 158	0.879	0.925	0.843	0.918	0.928	0.964
Observer set							
Benign vs malignant	20 / 80	0.975	0.960	0.963	0.988	0.850	0.944
Benign & grade group 1 vs grade group ≥ 2	32 / 68	0.905	0.870	0.899	0.912	0.781	0.806
Benign & grade group 1-2 vs grade group ≥ 3	48 / 52	0.870	0.850	0.794	0.962	0.729	0.946
External test set (pathologist 1 as reference standard)							
Benign vs malignant	12 / 233	0.987	0.976	0.987	0.987	0.750	0.750
Benign & grade group 1 vs grade group ≥ 2	91 / 154	0.870	0.829	0.833	0.909	0.692	0.818
Benign & grade group 1-2 vs grade group ≥ 3	119 / 126	0.856	0.841	0.804	0.913	0.765	0.892
External test set (pathologist 2 as reference standard)							
Benign vs malignant	10 / 235	0.991	0.984	0.996	0.987	0.900	0.750
Benign & grade group 1 vs grade group ≥ 2	40 / 205	0.903	0.841	0.919	0.888	0.600	0.511
Benign & grade group 1-2 vs grade group ≥ 3	69 / 176	0.883	0.824	0.845	0.926	0.565	0.750

3 Supplementary methods

3.1 Tumor detection system

A previously developed tumor detection system was applied[2] to outline tumor areas in our training set. To train this tumor detection system, a pathologist in training, supervised by an experienced uropathologist, outlined tumor regions in 100 prostate biopsies. Most of these biopsies were low grade: 52 benign, 11 grade group 1, 23 grade group 2, seven grade group 3, five grade group 4, and two biopsies grade group 5. The biopsies used for the development of the tumor detection system were independent of the biopsies used in the current study.

Patches were extracted from the 100 biopsies and used to train the system (pixel resolution of $1.92\mu m$). The tumor detection system achieved an AUC of 0.99 in discriminating benign from malignant biopsies on a separate test set of 75 biopsies.

After training the system was applied as a fully convolutional network to all biopsies of the current study. This procedure resulted in a rough outline of tumor regions.

3.2 Epithelium segmentation system

A previously developed epithelium segmentation system[3] was used to refine the tumor outlines generated by the tumor detection system. The epithelium segmentation system was developed using 102 prostatectomy tissue sections. The tissue sections were stained with H&E and subsequently restained with P63 and CK8/18 immunohistochemistry (IHC) markers to highlight epithelial structures. Each H&E and IHC pair were subsequently co-registered.

An initial deep learning system was trained on a subset of the IHC slides that were preprocessed with color deconvolution. This trained system was then applied to all IHC slides in the training set, forming a reference standard for the final system. This automated labeling method made sure that even poorly differentiated Gleason 5 areas were precisely annotated.

The final epithelium segmentation system was trained on the H&E slides using the automatically generated reference standard. A five-level-deep U-Net[4] was used as the network architecture with patches extracted at a pixel resolution of $0.98\mu m$. The system achieved a high segmentation performance (F1 score of 0.893) and was able to segment both intact glands and individual malignant epithelial cells.

3.3 Overview of deep learning system

Our deep learning system consisted of a U-Net[4] that was trained on randomly sampled patches extracted from the training set. After the automatic labeling process, the system could be trained on all biopsies, including those with mixed Gleason growth patterns. Additional patches were sampled from the hard-negative areas to improve the system's ability to distinguish tumor from benign tissue.

We followed the original U-Net architecture but experimented with different configurations. Given the importance of morphological features in Gleason grading, we focused on multiple depths of the U-Net, combined with different pixel spacings for the training data. Experimentation showed the best performance on the tuning set using a depth of six levels, sampled patches with a size of 1024×1024, and a pixel resolution of $0.96\,\mu m$. We added additional skip connections within each layer block and used up-sampling operations in the expansion path. Adam optimization was used with β_1 and β_2 set to 0.99, a learning rate of 0.0005 and a batch size of 8. The learning rate was halved after every 25 consecutive epochs without improvement on the tuning set. We stopped training after 75 epochs without improvement on the tuning set. Adding additional training examples from hard negative areas increased the performance of the network, especially in distinguishing between benign, inflammatory, and tumorous tissue.

The network was developed using Keras[5] and[6]. Data augmentation was used to increase the robustness of the network. The following augmentation procedures were used: flipping, rotating, scaling, color alterations (hue, saturation, brightness, and contrast), alterations in the H&E color space, additive noise, and Gaussian blurring.

3.4 Determining the Gleason grade group for a new specimen

Our deep learning system determines the Gleason grade group for a biopsy in two steps. First, our trained U-Net is applied to the scanned tissue of the biopsy. This procedure results in a label for each pixel of the image: background, stroma, benign epithelium, Gleason 3, Gleason 4, or Gleason 5. The frequency of each label can then be counted. Biopsies can differ vastly in size and in the amount of epithelial tissue that is present. To account for this difference, the values for the three Gleason growth patterns are normalized based on the sum of benign and malignant epithelial tissue. By normalizing, we obtain a volume estimate of the tumor that is independent of the size of the biopsy.

The volume percentages were used to assign a Gleason score and Gleason Grade group based on the guidelines for biopsies in clinical practice[7]. First, we determine whether a biopsy is malignant or benign. Based on the tuning set, we classify a biopsy as malignant if at least 10% of the epithelial tissue is predicted as cancer by the system. For malignant biopsies, we then determine the Gleason score. The growth pattern that has the largest volume is taken as the primary component. If there are other growth patterns present, with a volume of at least 7%, the most aggressive component is used as the secondary pattern.

The 7 and 10% cut-offs were determined automatically based on the tuning set. Note that this procedure differs between prostatectomies and biopsies. For prostatectomies, the secondary pattern is always the second-largest growth pattern, regardless of aggressiveness.

The predicted Gleason score is used to determine the grade group. A Gleason score 3+3 is mapped to group 1; Gleason score 3+4 is mapped to group 2; Gleason score 4+3 is mapped to group 3; Gleason scores 3+5, 4+4 and 5+3 are mapped to group 4; and higher scores are mapped to Gleason 5.

3.5 CycleGAN for style transformation and application to external data

We used a cycle-consistent generative adversarial network (CycleGAN)[8] system to facilitate stain transformation on the external dataset of tissue microarrays. In a CycleGAN setup, two separate networks are trained to perform a transformation from one stain to the other, while retaining the structural information of the tissue. We used a previously developed CycleGAN setup that was developed for the transformation of histopathological tissue[9]. The CycleGAN was adapted to learn the residual change, intead of reconstructing the full image.

To train the CycleGAN system, we sampled patches from the internal training set and used the full external dataset. Before inference with the Gleason deep learning system, we applied the CycleGAN network on the whole external dataset. The CycleGAN network was implemented in TensorFlow[6].

3.6 Determining Gleason score for the external dataset

The original paper of the external dataset[1] reports the quadratic kappa on Gleason score as the primary metric. The Gleason scores were determined using the standard for grading prostatectomies: the sum of the most and second most common growth patterns. We applied our algorithm to the test set of this paper and computed the quadratic kappa on Gleason score to allow a one-to-one comparison of our algorithm to the algorithm developed by Arvaniti et al. To account for the difference in the grading systems between prostatectomies and biopsies we adjusted the decision thresholds of the deep learning system using the training data set of 641 cores, resulting in a tumor threshold of 3% and a secondary pattern threshold of 1.5%.

3.7 Statistical analysis

To compare the performance of the deep with the external panel of pathologists, we performed multiple permutation tests. The permutation test was implemented as follows: For each case in the observer set, we obtained a list of predictions, one by the deep learning system and the remainder by the pathologists. In each iteration of the permutation test, for each case, we swapped the grade group prediction of the system with a random prediction from the list of predictions. After swapping the predictions, the test statistic was computed. Repeating this procedure 10.000 times resulted in a null distribution of the test statistic. The original test statistic was then compared to the null distribution, resulting in a two-tailed p-value.

We defined the main metric as the agreement with the consensus reference standard, measured using quadratic Cohen's kappa. For the analysis on agreement with the reference, we defined the test statistic as the difference between the kappa of the deep learning system and the medium kappa of the pathologists. The panel of pathologists was split into two groups, those with less than 15 years experience and those with more, and a permutation test was performed for both groups. The analysis of the ROC curves was done using the difference in F1-score as the metric. The decision threshold for the system was based on the point that maximized the AUC. The comparison of grade group accuracy was made using the difference in the accuracy of the system and the median accuracy of the pathologists with respect to the reference standard.

Supplementary references

1. Arvaniti, E. *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci. Reports* **8**, 1–11, DOI: 10.1038/s41598-018-30535-1 (2018).

2. Litjens, G. *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Nat. Sci. Reports* **6**, 26286, DOI: 10.1038/srep26286 (2016).

3. Bulten, W. *et al.* Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Sci. Reports* **9**, 1–7, DOI: 10.1038/s41598-018-37257-4 (2019).

4. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351 of *Lecture Notes in Computer Science*, 234–241 (2015).

5. Chollet, F. *et al.* Keras. https://github.com/keras-team/keras (2015).

6. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.

7. Epstein, J. I. *et al.* The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *The Am. J. Surg. Pathol.* **40**, 1, DOI: 10.1097/pas.0000000000000530 (2015).

8. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232 (2017).

9. de Bel, T., Hermsen, M., Kers, J., van der Laak, J. & Litjens, G. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In Cardoso, M. J. *et al.* (eds.) *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, vol. 102 of *Proceedings of Machine Learning Research*, 151–163 (PMLR, London, United Kingdom, 2019).