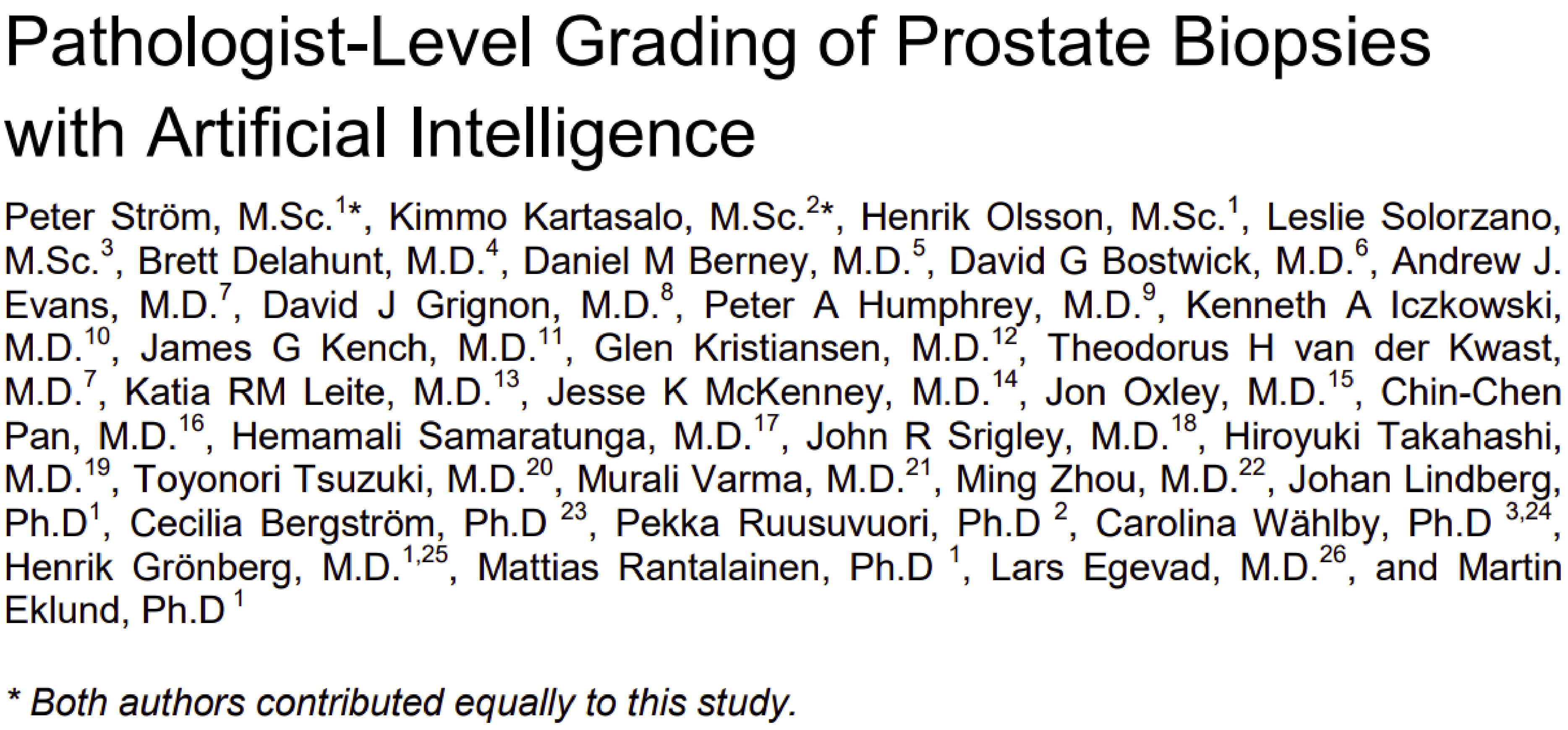


Pathologist-Level Grading of Prostate Biopsies with Artificial Intelligence

Peter Ström, M.Sc.[1*], Kimmo Kartasalo, M.Sc.[2*], Henrik Olsson, M.Sc.[1], Leslie Solorzano, M.Sc.[3], Brett Delahunt, M.D.[4], Daniel M Berney, M.D.[5], David G Bostwick, M.D.[6], Andrew J. Evans, M.D.[7], David J Grignon, M.D.[8], Peter A Humphrey, M.D.[9], Kenneth A Iczkowski, M.D.[10], James G Kench, M.D.[11], Glen Kristiansen, M.D.[12], Theodorus H van der Kwast, M.D.[7], Katia RM Leite, M.D.[13], Jesse K McKenney, M.D.[14], Jon Oxley, M.D.[15], Chin-Chen Pan, M.D.[16], Hemamali Samaratunga, M.D.[17], John R Srigley, M.D.[18], Hiroyuki Takahashi, M.D.[19], Toyonori Tsuzuki, M.D.[20], Murali Varma, M.D.[21], Ming Zhou, M.D.[22], Johan Lindberg, Ph.D[1], Cecilia Bergström, Ph.D [23], Pekka Ruusuvuori, Ph.D [2], Carolina Wählby, Ph.D [3,24], Henrik Grönberg, M.D.[1,25], Mattias Rantalainen, Ph.D [1], Lars Egevad, M.D.[26], and Martin Eklund, Ph.D [1]

** Both authors contributed equally to this study.*

P. Ström *et al.*, "Pathologist-Level Grading of Prostate Biopsies with Artificial Intelligence," *arXiv:1907.01368 [cs, eess]*, Jul. 2019 [Online]. Available: http://arxiv.org/abs/1907.01368.

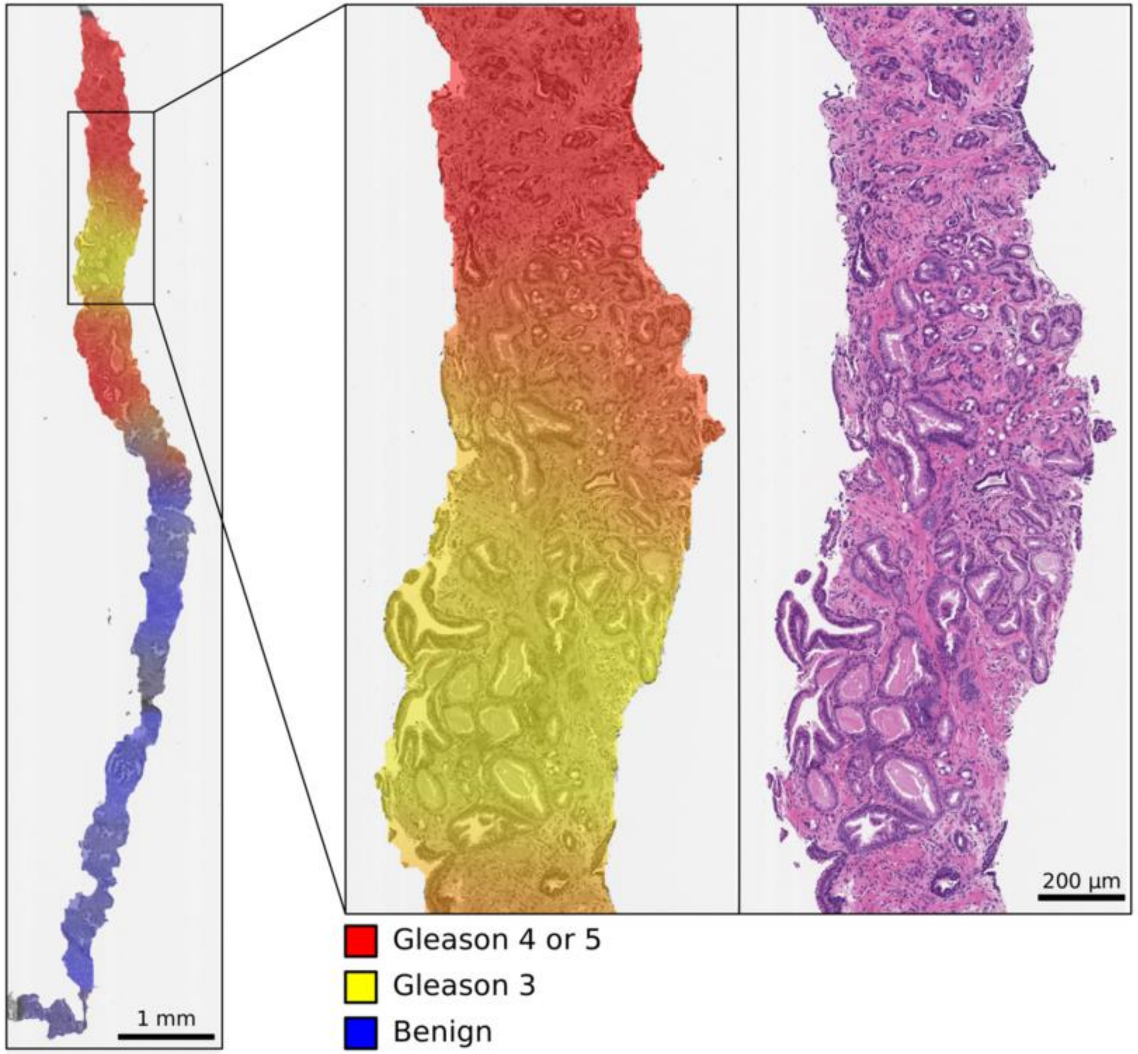


Figure 2: Color-coded visualization of cancer grades estimated by the AI. The colors represent the estimated probabilities for the presence of benign (blue), malignant low grade (Gleason 3, yellow) and malignant high grade (Gleason 4 or 5, red) tissue at different locations of the biopsy **(left)**. A magnified view of the AI output **(center)** and the corresponding H&E stained tissue **(right)** are shown for a region where an estimated transition between low- and high-grade morphology can be observed. This core from the test data was graded as ISUP 3 (GS 4+3) by the study pathologist.

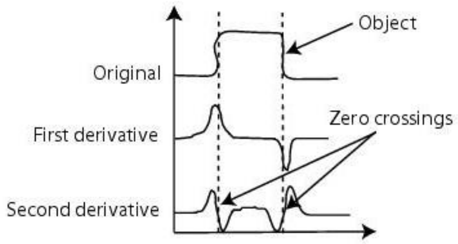
Image pre-processing

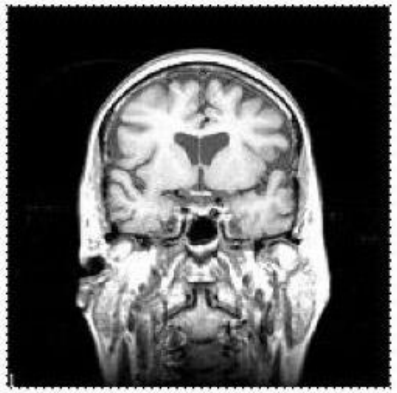
1. Segmentation of tissue
2. Segmentation of pen marks
3. Digitization of annotations (Extraction of digital pixel-wise annotation)
4. Extracting partially overlapping patches from the whole slide images (Patch dimensions: 598 x 598 pixels (approx. 540 x 540 µm) at a resolution corresponding to 10X magnification (pixel size approx. 0.90 µm))

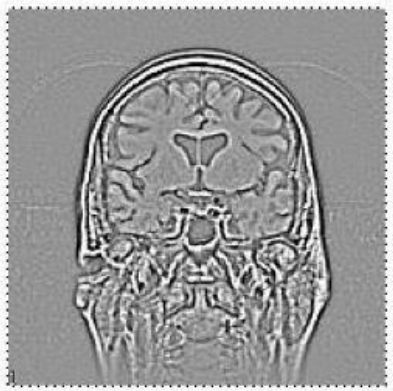
Segmentation of tissue

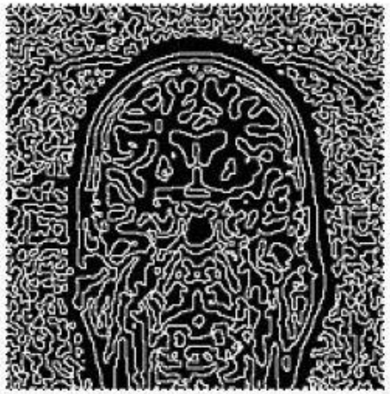
1. Read images downsampled by a factor 16
2. Conversion from RGB to grayscale (0.2989 x R + 0.5870 x G + 0.1140 x B)
3. 2D Laplacian filtering
4. The absolute magnitude of the resulting response was thresholded using Otsu’s method
5. Morphological closing with a disk-shaped structuring element having a radius of 50 µm (filling of holes and removal of objects having an area smaller than 100 000 µm 2)
6. Any remaining pen marks were removed based on their color by performing the HSV transform and excluding any objects whose mean hue was less than 0.7
7. Rescale back to full resolution using nearest neighbour interpolation

Laplacian filtering and Otsu's method




(A) Original MR image


(B) Laplacian results


(C) Extraction of the zero crossing of the Laplacian (object edges)

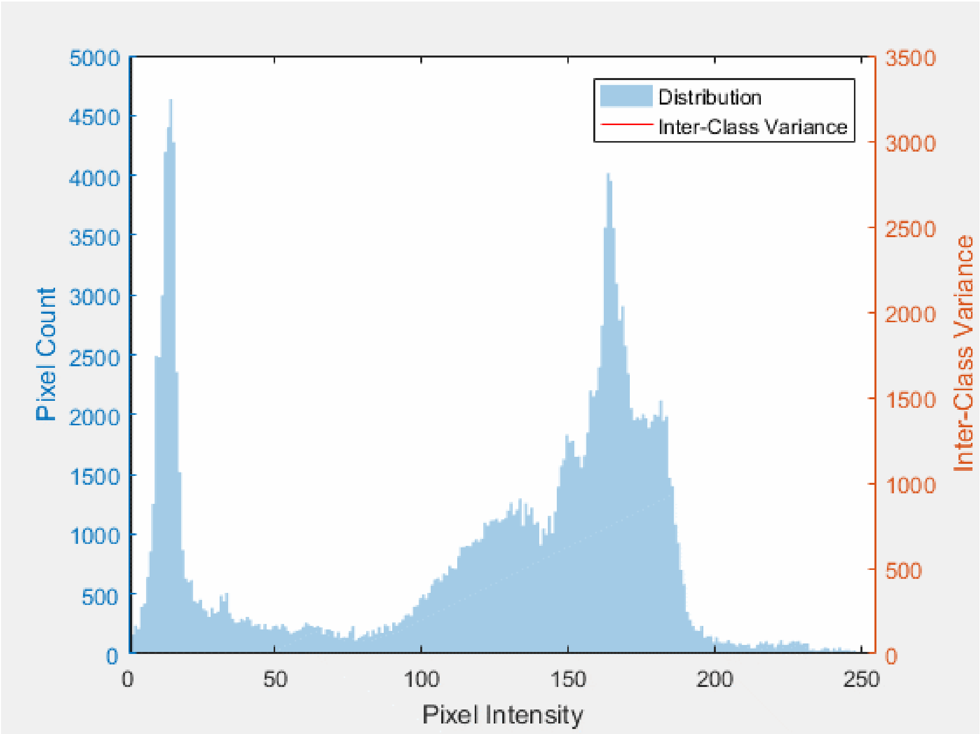


Image pre-processing

1. Segmentation of tissue
2. Segmentation of pen marks
3. Digitization of annotations (Extraction of digital pixel-wise annotation)
4. Extracting partially overlapping patches from the whole slide images (Patch dimensions: 598 x 598 pixels (approx. 540 x 540 µm) at a resolution corresponding to 10X magnification (pixel size approx. 0.90 µm))

Architecture

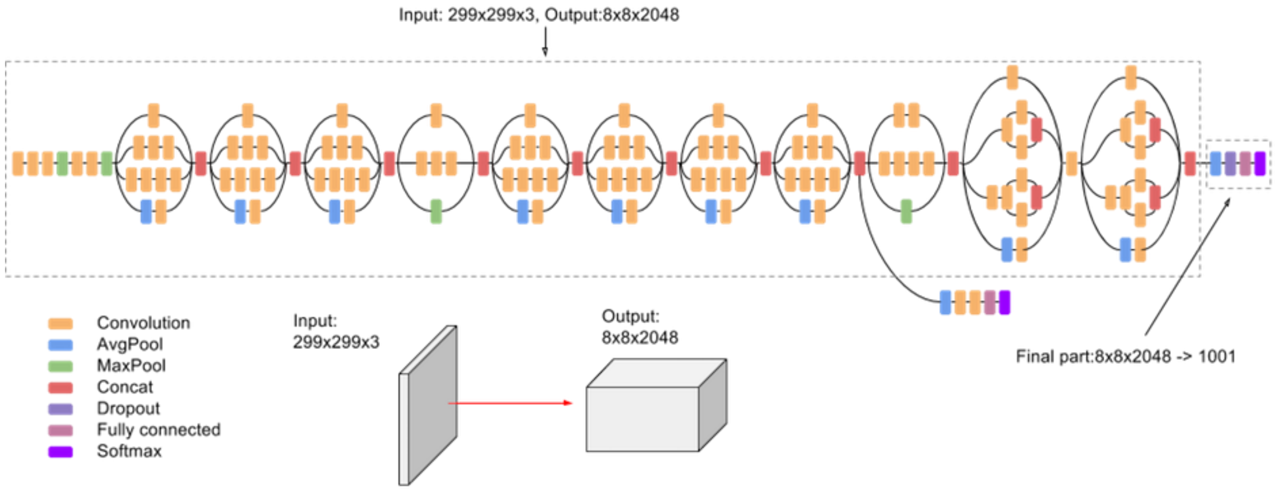
Two DNN ensembles, each consisting of 30 Inception V3 models pretrained on ImageNet

- First ensemble: binary classification – benign or malignant
- Second ensemble: classification into Gleason patterns 3 to 5 (only one Gleason pattern per patch)

The probabilities for the Gleason pattern are obtained from the DNN ensembles.

Boosted trees:

- Input: aggregating features from the patch-wise probabilities predicted by each DNN for each core
- Output: The clinical assessment of Gleason score and cancer length.



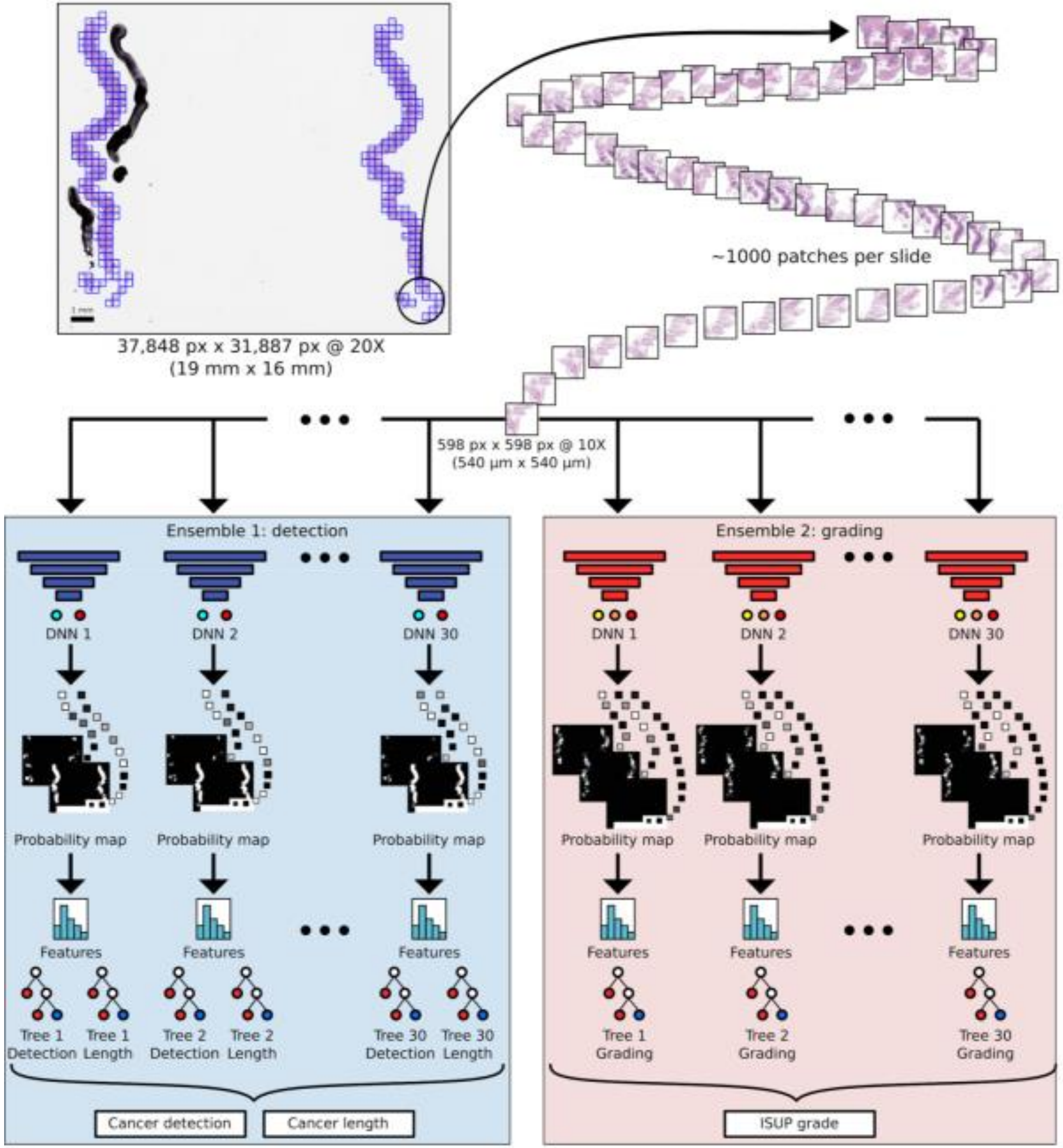


Figure S2: Overview of the artificial intelligence system. The tissue region in the input WSI is split into patches **(top)**. The patches are fed as input to a detection ensemble of 30 DNNs for discriminating between benign and malignant patches **(left box; top row)**, and to a grading ensemble of 30 DNNs for classifying patches into Gleason grades 3, 4 and 5 **(right box; top row)**. In the patch-level prediction phase **(both boxes; middle row)**, each trained DNN outputs a vector of class-wise probabilities for each input patch. The class-wise probabilities, indicated here with squares where the grayscale intensity corresponds to probability value, are mapped back to the locations of the corresponding patches in the WSI to construct a probability map for each class. The maps are summarized into features, which are used as inputs to train ensembles of boosted trees to predict cancer presence and extent **(left box; bottom row)** and grade **(right box; bottom row)** for entire WSIs. In the WSI-level prediction phase, outputs from the 30 boosted trees in each ensemble are averaged, and the final classification of each WSI is assigned to the class associated with the highest average probability.

Training

Due to the **lack of annotations** indicating where each Gleason pattern is located on slides with multiple patterns (e.g. 3+4), we **only used** cancerous patches from slides with a single Gleason pattern (e.g. 3+3) for training.

a) As an alternative to discarding slides with multiple Gleason patterns when training the second stage model, we evaluated an iterative approach. This involved:

   1) training a model only on slides featuring a single Gleason pattern,
   2) applying this model to predict the patch-level Gleason grades for slides featuring multiple patterns and
   3) either fine-tuning the model further or training a new model from scratch using the predicted grades as additional training labels.

b) Moreover, we experimented with using the primary Gleason grade (e.g. 3 for a 3+4 slide) as the label for all patches from a slide, or randomly choosing the Gleason grade for each patch from the two slide-level grades.

Summary: None of these approaches improved classification performance compared to only using single Gleason slides as training data.

In order to compensate for the considerably **imbalanced distribution of different classes** in the training data, which can be detrimental for CNN models, we performed class balancing via subsampling before each ‘epoch’. Specifically, we **always included all training patches representing the rarest class**, and r**andomly sampled the same amount of patches from all other classes**. This results in a uniform class distribution without duplicated examples. The model was then trained until this set of patches was exhausted, and the random sampling step was repeated with replacement for the next ‘epoch’.

To improve generalization and to obtain rotational invariance, each patch was randomly rotated by either **0°, 90°, 180° or 270°**, and then **flipped vertically** with a probability of 50% before being fed to the CNN.

Evaluation metrics

Cancer detection:

Receiver Operating Characteristic (ROC) curve and the Area Under the ROC Curve (AUC)

Gleason grading:

To evaluate how well the AI agreed with the pathologists, they calculated all pair-wise kappas and summarized the average for each of the raters. In addition, they estimated the kappa with a grouping of the Gleason scores in ISUP grades (grade groups) 1, 2-3 and 4-5.

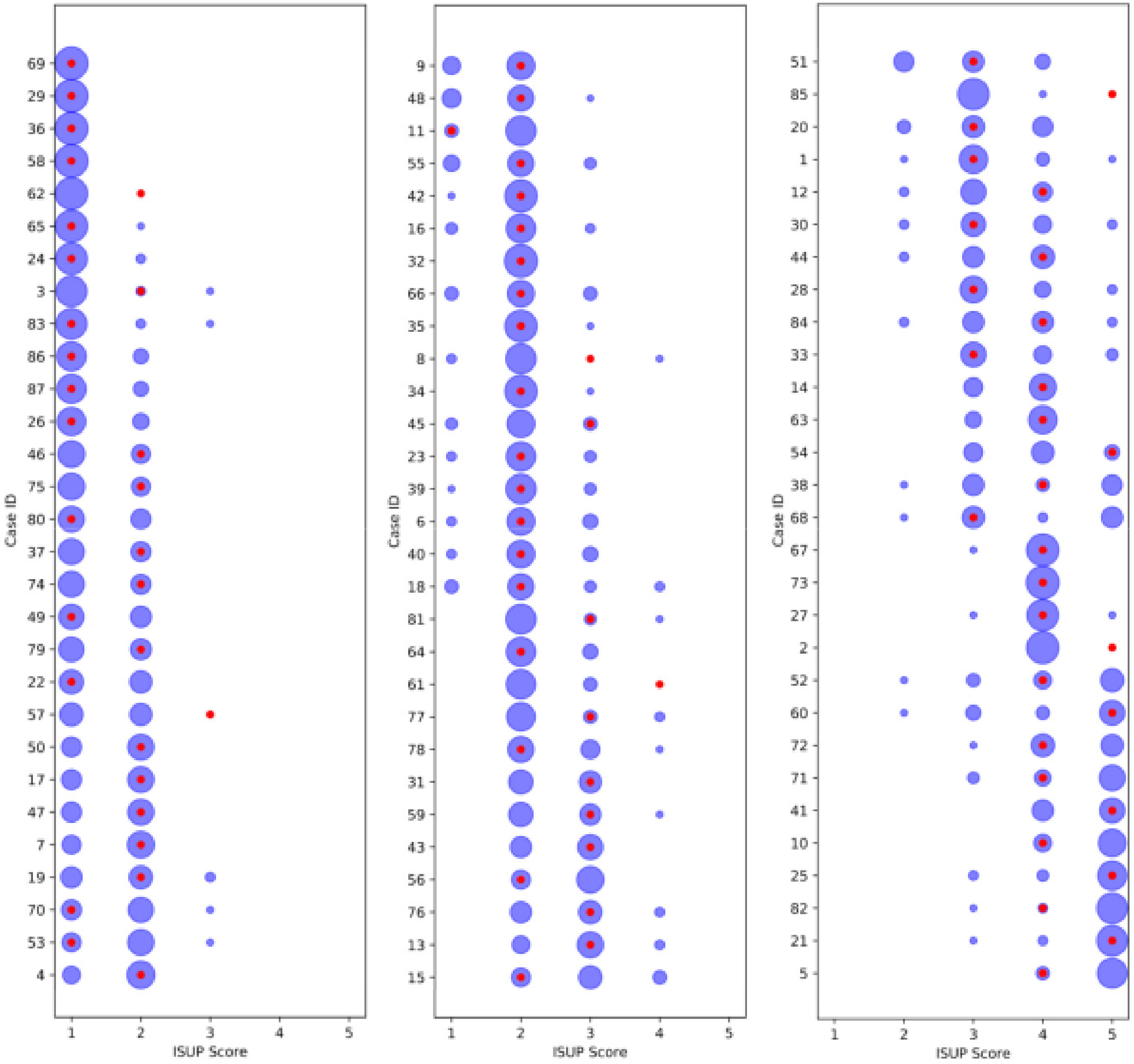
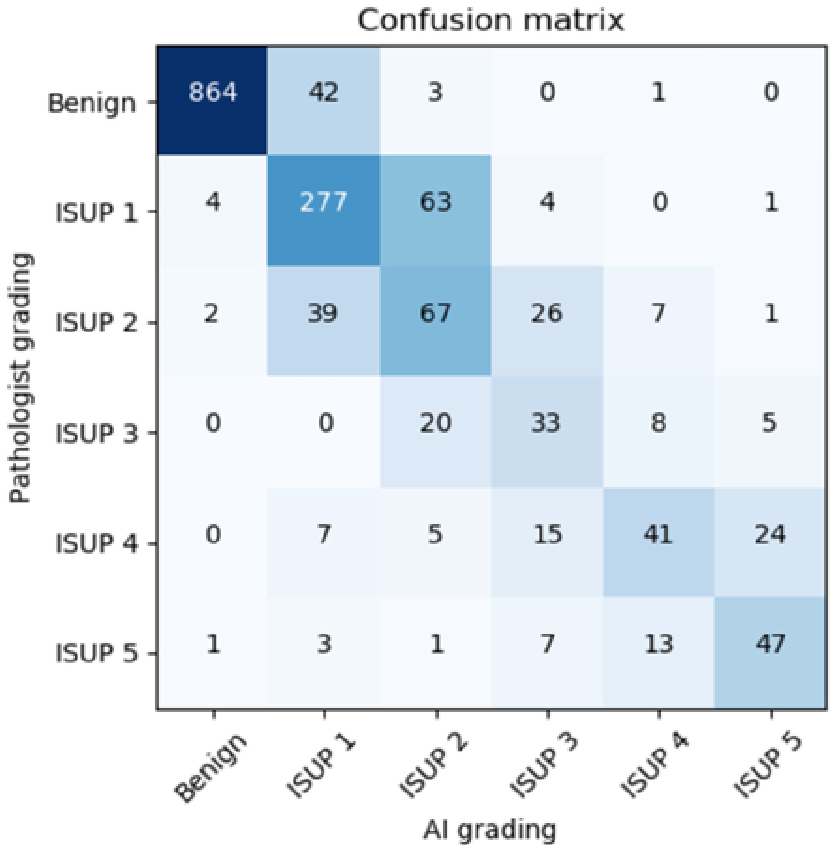


Figure S3: Grading performance relative to ISUP expert panel on Imagebase. The distribution of ISUP scores given by the 23 pathologists from the ISUP expert panel and the AI for each of the 87 case IDs in Imagebase. Each row corresponds to one case, and the cases are organized into three plots according to average ISUP score increasing from left to right, and from top to bottom. The areas of the blue circles represent the proportion of pathologists who voted for a specific ISUP score (x-axis). The red dot indicates the ISUP score given by the AI. Example: in the last row (bottom-right; case ID 5) most pathologists voted ISUP 5 and a minority voted ISUP 4; the red dot indicates that the AI voted ISUP 4.



	Benign	ISUP 1	ISUP 2	ISUP 3	ISUP 4	ISUP 5
Benign	864	42	3	0	1	0
ISUP 1	4	277	63	4	0	1
ISUP 2	2	39	67	26	7	1
ISUP 3	0	0	20	33	8	5
ISUP 4	0	7	5	15	41	24
ISUP 5	1	3	1	7	13	47

Figure S4: Grading performance on independent test data. A confusion matrix on the independent test data of 1631 slides. The pathologist's (L.E.) grading is shown on the y-axis and the AI's grading on the x-axis. Cohen's kappa with linear weights was 0.83 when considering all cases, and 0.70 when only considering the cases indicated as positive by the pathologist.

Table S2: Comparison of CNN architectures. Cancer detection (AUC), cancer length estimation (Correlation) and grading (Cohen's kappa) performance of ensembles representing different CNN architectures, estimated using cross-validation on the training data. The highest value for each metric is highlighted in bold italic.

Architecture (epochs)	**AUC**	**Correlation**	**Cohen's kappa**
Inception V3 (20)	0.984	***0.943***	***0.64***
Inception V3 (30)	***0.987***	0.939	0.63
Inception V3 (40)	0.984	0.935	0.62
ResNet-50 (20)	0.983	0.933	0.57
ResNet-50 (30)	0.980	0.935	0.60
ResNet-50 (40)	0.982	0.932	0.55
Inception-ResNet V2 (20)	0.983	0.939	0.61
Inception-ResNet V2 (30)	0.984	0.939	0.62
Xception (20)	0.985	0.937	0.58
Xception (30)	0.985	0.939	0.58

Table S4: Effect of training epochs on cancer detection and length estimation performance. Cancer detection (AUC) and length estimation (Correlation) performance as a function of training epochs for an ensemble of 5 Inception V3 models, estimated using cross-validation on the training data.

	10 epochs	15 epochs	20 epochs
AUC	0.991	0.993	0.992
Correlation	0.947	0.949	0.948

Table S5: Effect of training epochs on cancer grading performance. Cancer grading performance (Cohen’s kappa) as a function of training epochs for an ensemble of 5 Inception V3 models, estimated using cross-validation on the training data.

	40 epochs	60 epochs	80 epochs
Cohen’s kappa	0.66	0.67	0.66

None of these studies achieved expert uro-pathologist level consistency in Gleason grading, estimated tumor burden or investigated the reproducibility of grading on needle biopsies, which are utilized for diagnostics in virtually every pathology laboratory worldwide.

The main limitation of this study is the lack of exact pixel-wise annotations, since the annotations may highlight regions that include a mixture of benign and malignant glands of different grades.

Comments

- Horrible order of the article (short article version, figures, long article version, figures, tables)
- Pre-processing Matlab, training Python (Keras)
- Model training took approximately 5 or 3 days per one first stage or second stage CNN, respectively.