Patterns, Volume *3*

Supplemental information

An AI-assisted tool for efficient prostate cancer diagnosis in low-grade and low-volume cases

Mustafa Umit Oner, Mei Ying Ng, Danilo Medina Giron, Cecilia Ee Chen Xi, Louis Ang Yuan Xiang, Malay Singh, Weimiao Yu, Wing-Kin Sung, Chin Fong Wong, and Hwee Kuan Lee

SUPPLEMENTAL ITEMS

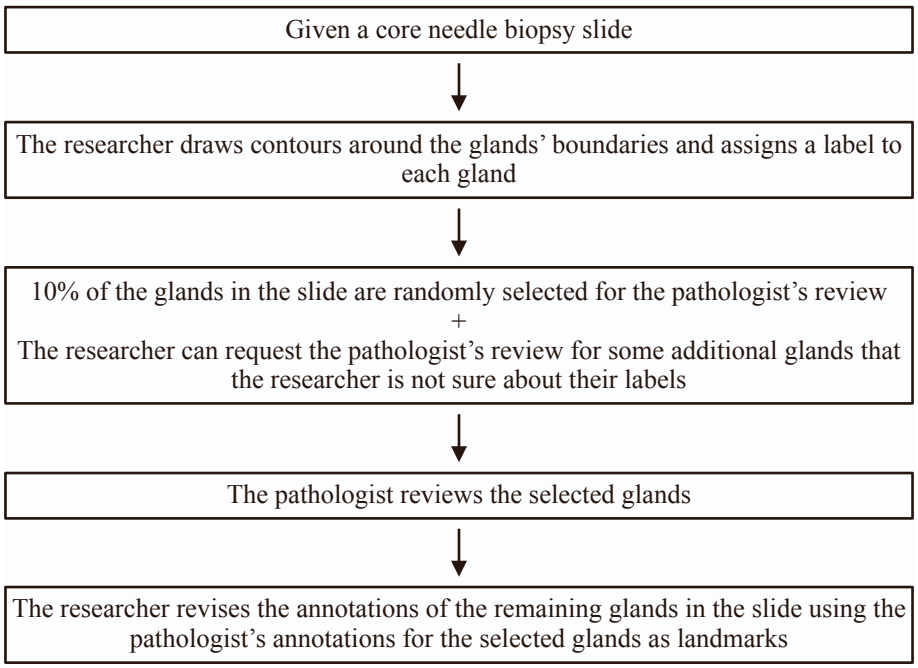
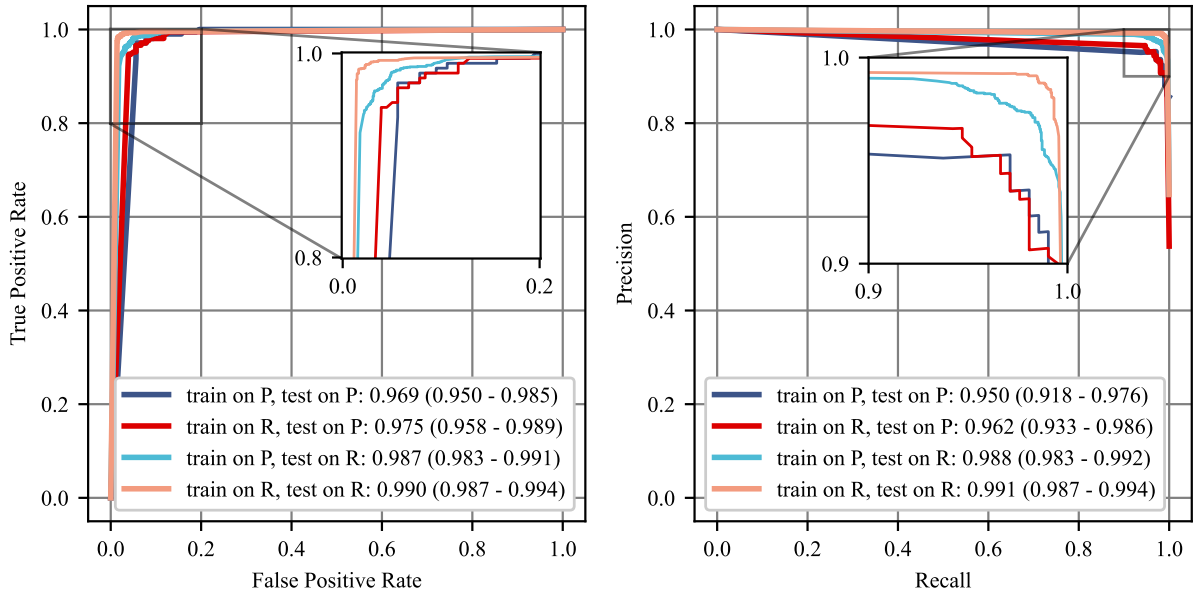


Figure S1: The workflow for the annotation of core needle biopsy slides. Related to Figure 1A.

Table S1: The number of slides in training, validation, and test sets in the PANDA dataset. The values in parentheses show the percentages. Related to Table 1.

	Radboud				Karolinska				PANDA			
	Train	Valid	Test	Total	Train	Valid	Test	Total	Train	Valid	Test	Total
No. of slides	3,021	1,007	1,032	5,060	3,258	1,086	1,108	5,452	6,279	2,093	2,140	10,512
Non-tumor	567	189	192	948 (19)	1,152	384	388	1,924 (35)	1,719	573	580	2,872 (27)
Tumor-containing	2,454	818	840	4,112 (81)	2,106	702	720	3,528 (65)	4,560	1,520	1,560	7,640 (73)
3+3	480	160	162	802 (20)	1086	362	364	1812 (51)	1,566	522	526	2,614 (34)
3+4	402	134	137	673 (16)	399	133	134	666 (19)	801	267	271	1,339 (17)
4+3	543	181	185	909 (22)	189	63	66	318 (9)	732	244	251	1,227 (16)
4+4	393	131	132	656 (16)	279	93	94	466 (13)	672	224	226	1,122 (15)
3+5	39	13	15	67 (2)	6	2	5	13 (0)	45	15	20	80 (1)
5+3	24	8	9	41 (1)	0	0	2	2 (0)	24	8	11	43 (1)
4+5	378	126	130	634 (15)	123	41	44	208 (6)	501	167	174	842 (11)
5+4	132	44	45	221 (5)	15	5	7	27 (1)	147	49	52	248 (3)
5+5	63	21	25	109 (3)	9	3	4	16 (1)	72	24	29	125 (2)



(a) AUROC **(b)** AP

Figure S2: Performance evaluation of the models trained on annotations by the pathologist (P) and the researcher (R). (a) Area under receiver operating characteristics curve (AUROC) and (b) average precision (AP) calculated over precision vs. recall curve together with 95% confidence intervals (obtained using the percentile bootstrap method[1]) are presented. Note that models were trained and tested on the training set and test set of the gland classification dataset, respectively.

Table S2: Performance of our algorithms in prostate cancer detection. The PANDA model was a three-resolution classification model trained on the training set of the PANDA dataset. The SG pipeline consisted of gland segmentation Mask R-CNN model and four-resolution gland classification model which were trained on training sets of SG gland segmentation and classification datasets, respectively. Related to Table 3.

Model	Dataset	# of slides/parts (B: Benign, M: Malignant)	AUROC (95% CI)
PANDA Model	PANDA test set (internal)	2140 CNB slides (B=580, M=1560)	0.972 (0.965 - 0.978)
PANDA Model	SG dataset (external)	280 CNB parts (B=179, M=81)	0.992 (0.985 - 0.997)
PANDA Model	SG test set (external)*	81 CNB parts (B=50, M=31)	0.980 (0.953 - 0.997)
SG pipeline	SG test set (internal)	81 CNB parts (B=50, M=31)	0.997 (0.987 - 1.000)

CNB: Core Needle Biopsy.
*The SG test set is a subset of the SG (gland classification) dataset.

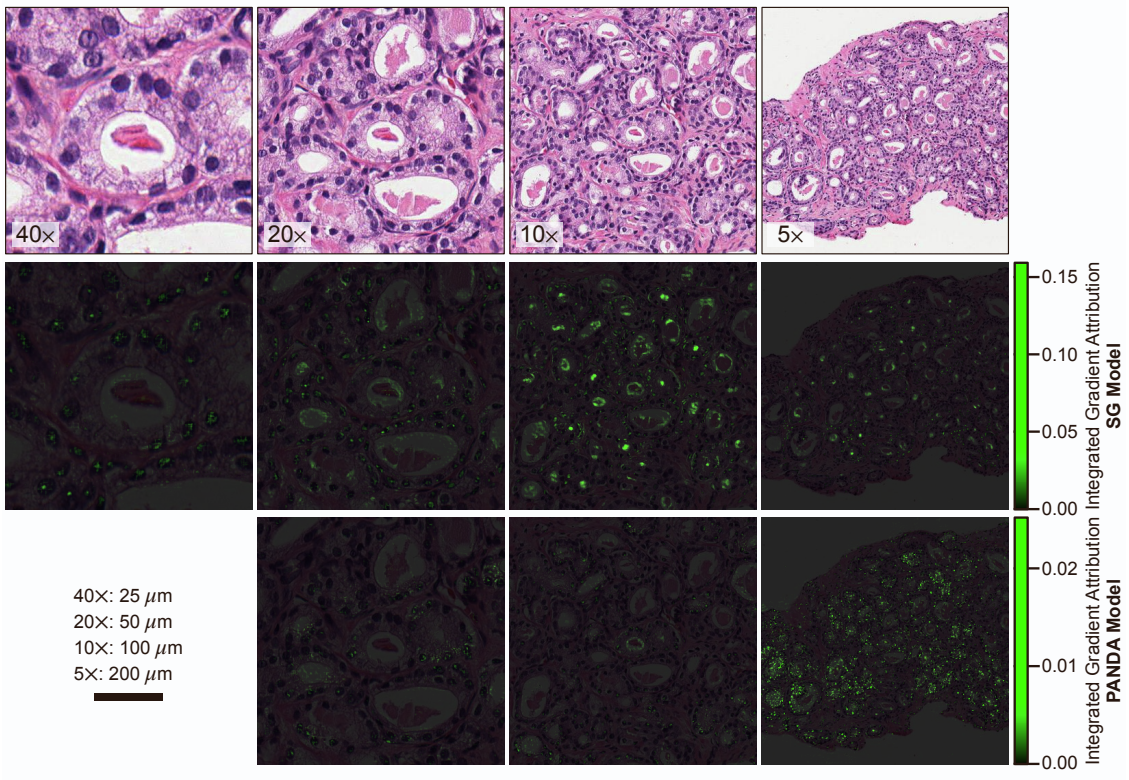


Figure S3: Post-hoc analysis on a malignant sample using the trained SG and PANDA models. Attribution maps obtained using integrated gradients[2] with blurred images as baselines are presented for a malignant sample in the test set of the SG gland classification dataset. This sample was predicted correctly by the SG and PANDA models. Related to Figure 3A.

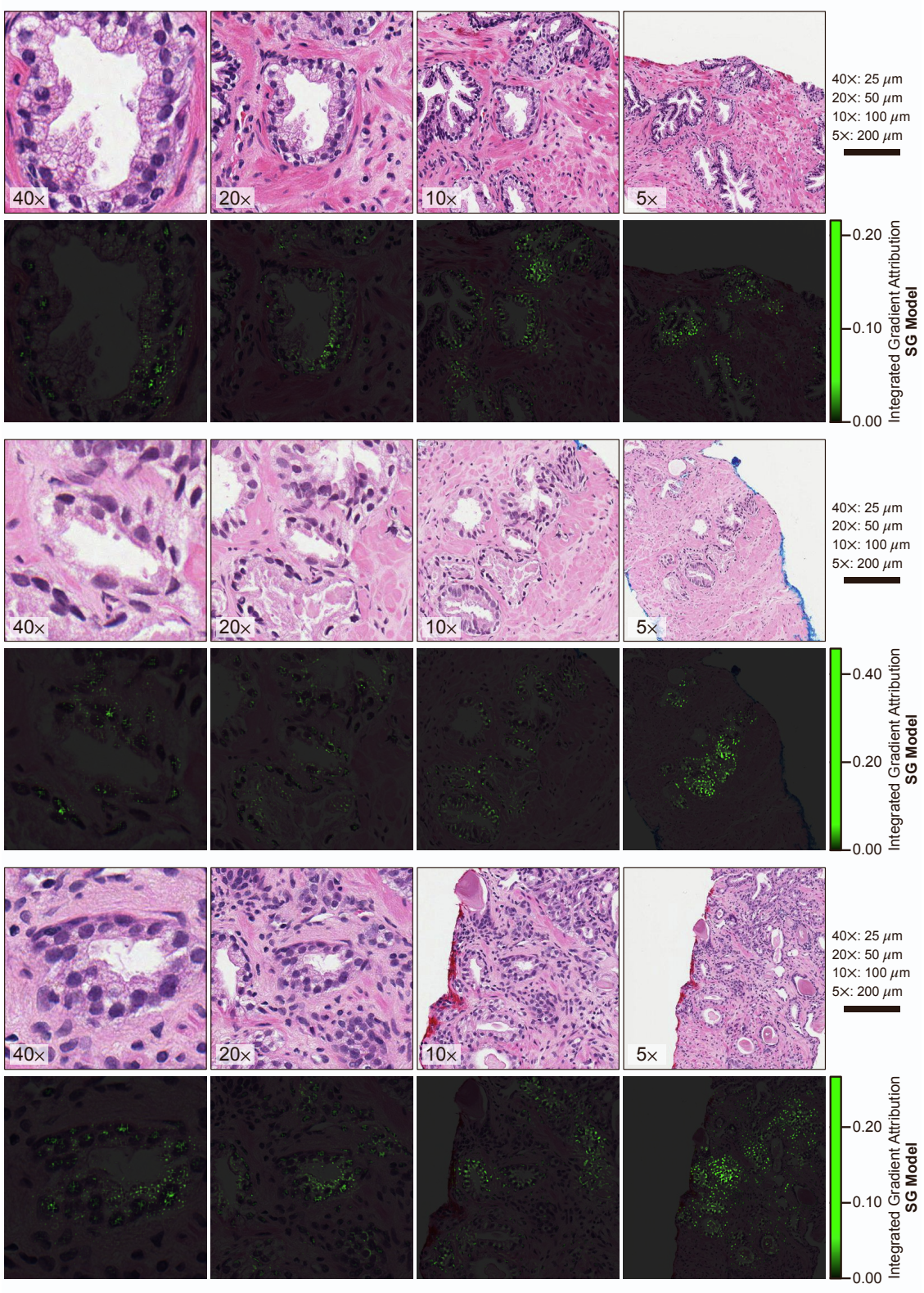


Figure S4: Post-hoc analysis on three benign samples using the trained SG model. Attribution maps obtained using integrated gradients[2] with white images as baselines are presented for three benign samples in the test set of the SG gland classification dataset. These samples were predicted correctly by the four-resolution model trained on the SG model. Related to Figure 3A.

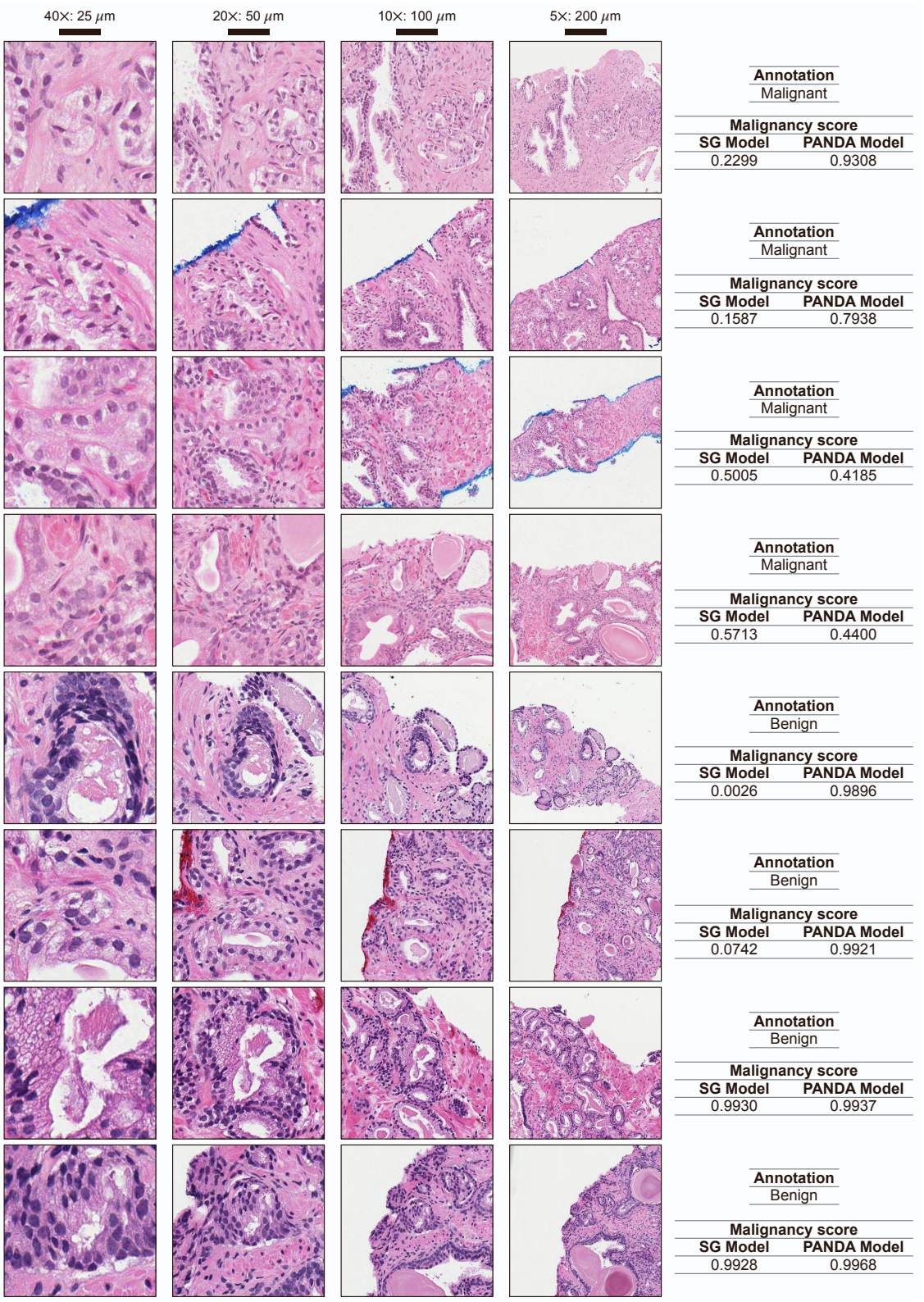


Figure S5: Example predictions by the trained SG and PANDA models. Example patches from the test set of the SG gland classification dataset together with annotations by the pathologist and predicted malignancy scores by the trained SG and PANDA models are presented. Scale bars are shown in black. Related to Figure 3B.

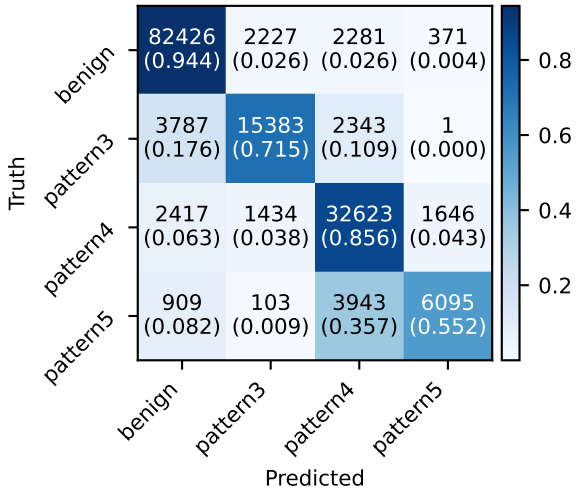


Figure S6: Confusion matrix for patch-level Gleason pattern predictions. Gleason pattern predictions for all patches within slides in the PANDA Radboud test set were obtained from the trained multi-resolution Gleason pattern prediction model. The values in parentheses show the row-wise percentages.

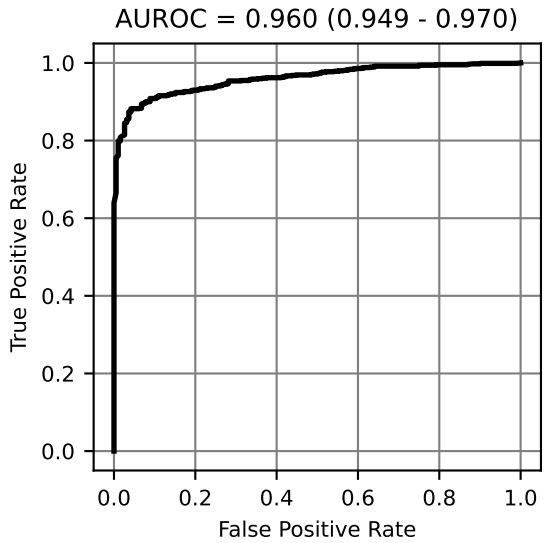


Figure S7: Benign vs. malignant slide classification using multi-resolution Gleason pattern prediction model. Gleason pattern predictions for all patches within slides in the PANDA Radboud test set were obtained from the trained multi-resolution Gleason pattern prediction model. From patch predictions of a slide, a malignancy score was obtained for the slide. Then, a ROC curve analysis was conducted for benign vs. malignant slide classification over malignancy scores. An AUROC value of 0.960 (95% CI:0.949 - 0.970) was obtained.

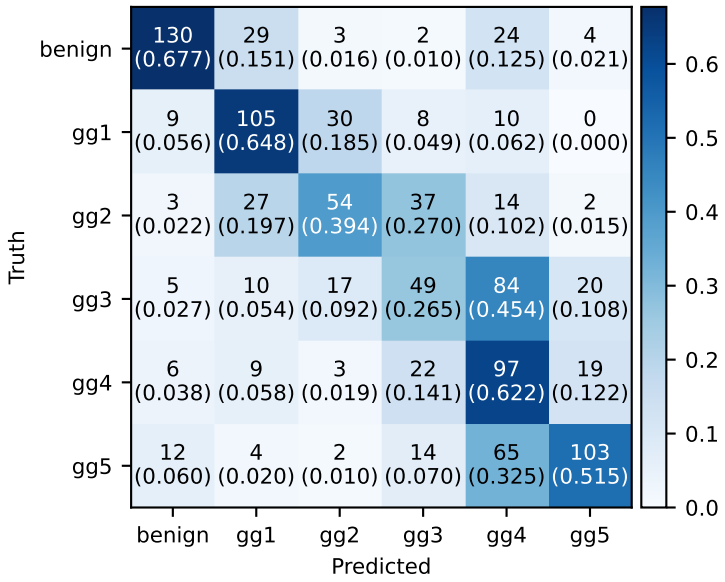


Figure S8: Confusion matrix for slide-level Gleason grade group predictions. Gleason pattern predictions for all patches within a slide in the PANDA Radboud test set were obtained from the trained multi-resolution Gleason pattern prediction model. Then, grade group (gg) predictions were obtained based on the proportion of predicted Gleason patterns within a slide. A quadratically weighted Cohen's κ value of 0.707 (95% CI:0.665 - 0.748) between the slide labels and predicted grade groups was obtained. The values in parentheses show the row-wise percentages.

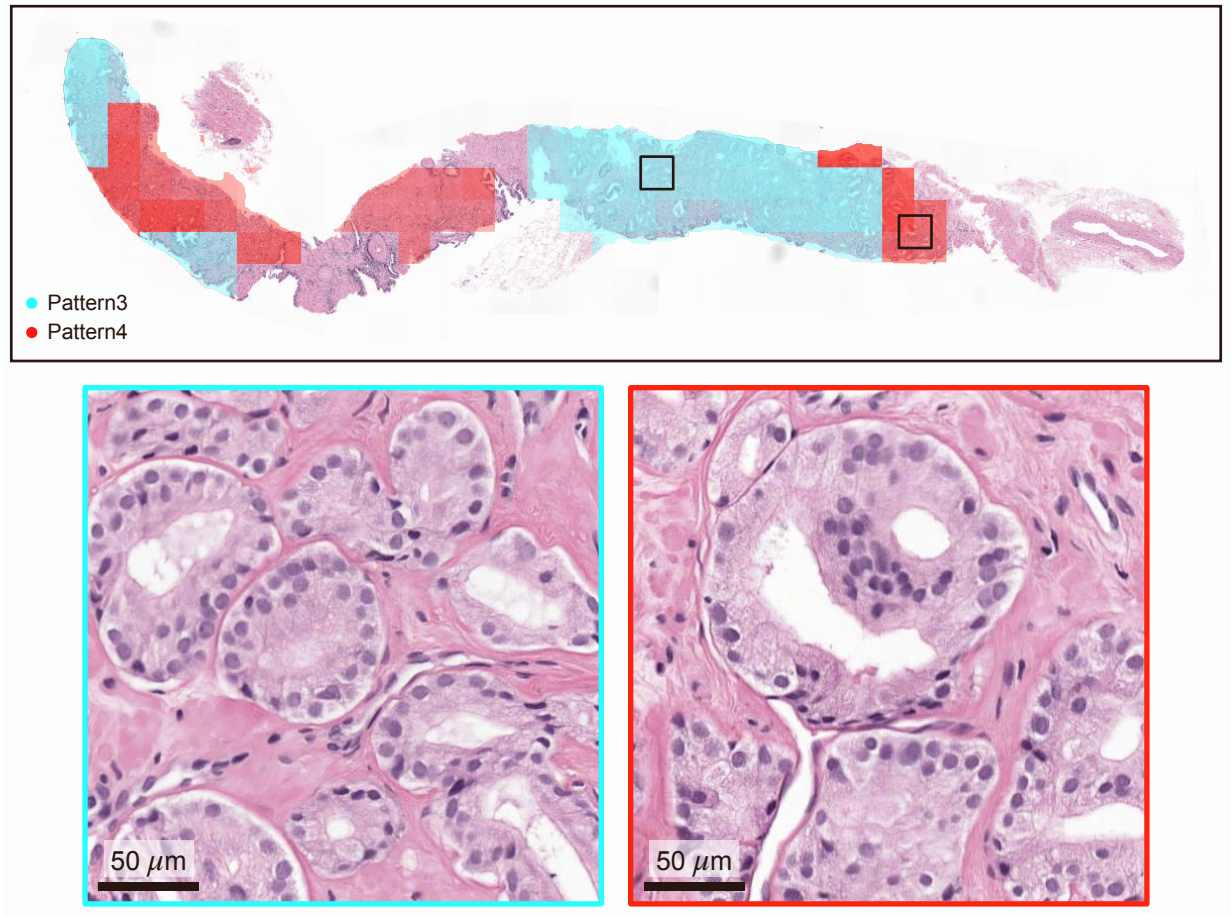


Figure S9: Color-coded Gleason pattern heatmap on a GS 3+4 patient's slide in the Radboud test set. Predicted patterns by the multi-resolution model are color-coded and overlayed on the original slide. The calculated percentages for Gleason patterns are: *benign* 43%, *pattern3* 35%, and *pattern4* 22%. Gradients in color codes indicate prediction scores for the corresponding pattern. Besides, two high-resolution patches are presented from two different pattern regions. Border color of a patch indicates the predicted pattern for the patch.

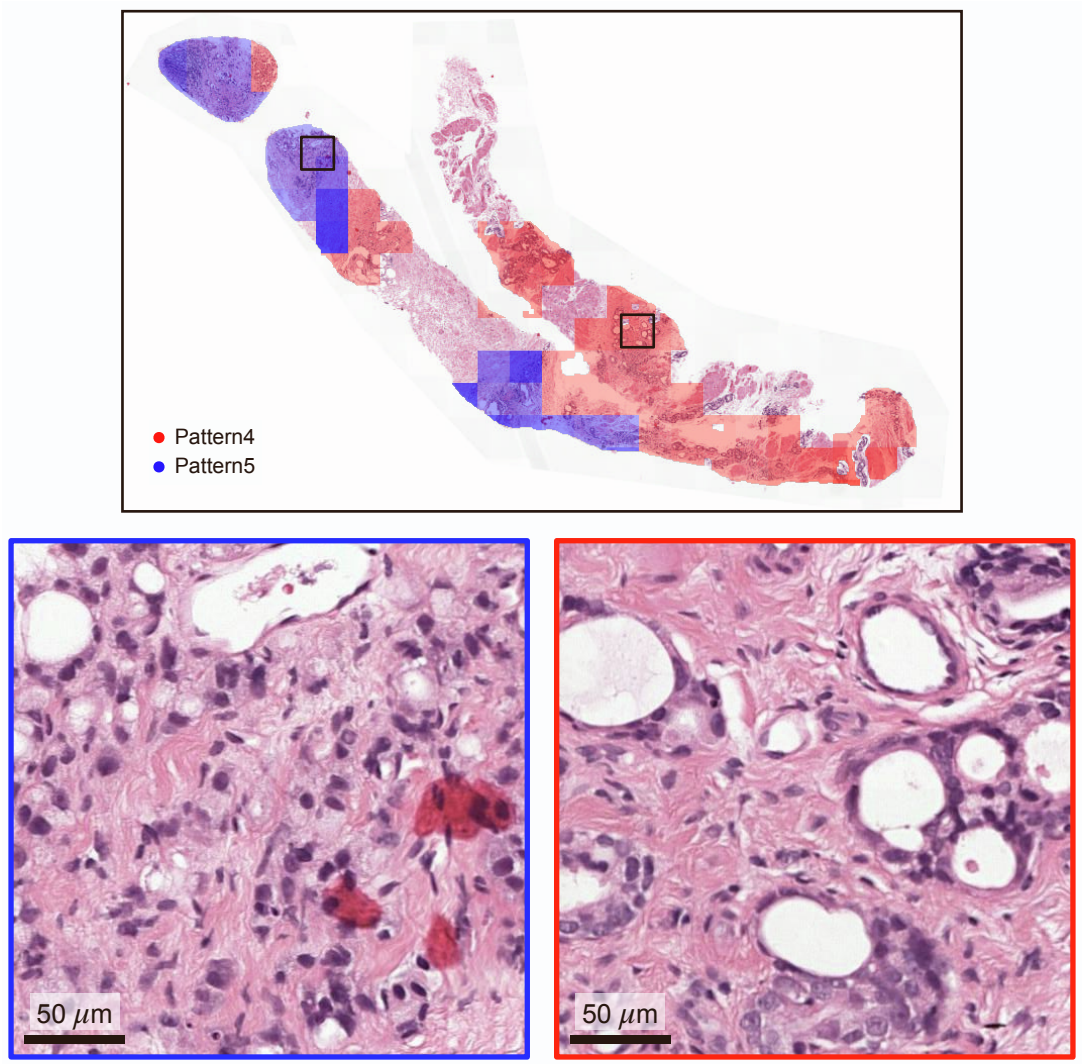


Figure S10: Color-coded Gleason pattern heatmap on a GS 4+5 patient's slide in the Radboud test set. Predicted patterns by the multi-resolution model are color-coded and overlayed on the original slide. The calculated percentages for Gleason patterns are: *benign* 38%, *pattern4* 39%, and *pattern5* 23%. Gradients in color codes indicate prediction scores for the corresponding pattern. Besides, two high-resolution patches are presented from two different pattern regions. Border color of a patch indicates the predicted pattern for the patch.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Post-processing predicted masks

Predicted gray-scale mask for each instance was thresholded at 0.5 to obtain a binary mask. When thresholding resulted in multiple contours, the largest contour was used to represent a segmented instance (gland). Holes in the binary mask, if any, were then filled. Predictions at tissue boundaries extending to the background were excluded.

Multiple binary masks corresponding to the same gland were observed due to the use of overlapping patches and detection of incomplete glands at the boundary. Thus, the masks were processed to remove redundant ones as follows:

1. When the intersection-over-union (IoU) or intersection over minimum area between two predicted masks exceeded the threshold of 0.3, the mask with the lower prediction score was discarded.
2. A mask that intersected with two or more masks was excluded.
3. Finally, two binary masks that had an IoU exceeding 0.3 were merged in an iterative manner starting with the pair of masks that had the greatest IoU.

Multi-resolution Gleason pattern classification model

We modified our multi-resolution benign vs. malignant patch classification model into Gleason pattern classification model with four classes: benign, pattern3, pattern4, and pattern5. This was a three-resolution model accepting $20\times$, $10\times$, and $5\times$ patches at the input and predicting Gleason pattern at the output.

Training of the model

The model was trained end-to-end from scratch using Adam optimizer for 2152 iterations. The model was trained on the training set of the Radboud dataset (Table S1), and performance on the validation set was tracked for early stopping. The learning rate was initially set to $5e-4$ and reduced to $5e-5$ at the end of iteration 1506, where the validation set performance was saturated. A weight decay of $5e-5$ was also used for regularization. Batch size was 16.

Benign vs. malignant slide classification

Predictions for all patches within a slide were obtained from the trained Gleason pattern classification model. Then, a four-channel heatmap for the slide by mapping the obtained class scores into corresponding patch locations. To eliminate outliers, a 2×2 moving average filter was applied on the heatmap.

To conduct a benign vs. malignant classification study, we obtained a malignancy score for each slide. Malignant channels (pattern3, pattern4, and pattern5) in the heatmap were aggregated by summing them up. The maximum score in the resulting channel was used as the slide's malignancy score. Finally, a receiver operating characteristics curve analysis was conducted (Figure S7).

Gleason grade group prediction

The pattern with the highest score at a point in the smoothed heatmap was assigned as that point's Gleason pattern prediction (Figure S9 and S10). Based on these predictions, percentages of patterns within a slide were calculated. Finally, a slide's grade group was obtained using Algorithm S1.

```python
import numpy as np

gs_to_gg_dict = { '0+0':0, '3+3':1, '3+4':2, '4+3':3, '4+4':4,
                  '3+5':4, '5+3':4, '4+5':5, '5+4':5, '5+5':5 }

def get_gg(percent_patterns):
  sorting_indices = np.argsort(percent_patterns)

  # how many patterns are there
  pattern_count = np.sum(percent_patterns>0)

  if pattern_count == 0:
    first_pattern = 0
    second_pattern = 0
  else:
    first_pattern = sorting_indices[-1] + 3

    if temp_percentages[sorting_indices[-2]]<0.05:
      second_pattern = first_pattern
    else:
      second_pattern = sorting_indices[-2] + 3

      # for biopsy slides, the highest grade is reported as 2nd pattern if it is > 5%
      if sorting_indices[0]==2 and temp_percentages[sorting_indices[0]]>0.05:
        second_pattern = 5

  gs = '{}+{}'.format(first_pattern,second_pattern)
  gg = gs_to_gg_dict[gs]

  return gg
```

Algorithm S1: Obtaining grade groups based on percentage of Gleason patterns within a slide.

SUPPLEMENTAL REFERENCES

[1] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics, pages 569–593. Springer.

[2] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In International conference on machine learning, pages 3319–3328. PMLR.