

Supplementary Online Content

Nagpal K, Foote D, Tan F, et al. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA Oncol*. Published online July 23, 2020. doi:10.1001/jamaoncol.2020.2485

eMethods

eFigure 1. Overview of the Reviews Provided by Expert Subspecialists for the Validation Set and an Overview of the Deep Learning System

eFigure 2. Example Overcalls by the DLS on Biopsies Without Tumor

eFigure 3. Validation Set Performance for Gleason Grading and Tumor Detection

eFigure 4. Evaluation of Inter-Subspecialist Agreement

eFigure 5. Comparison of the Deep Learning System (DLS) With Individual Subspecialists on Tumor-Containing Slides

eFigure 6. Sensitivity of the DLS's Gleason Grade Group Classification Accuracy to Differences in Gleason Pattern (GP) Thresholds for GP 3 vs 4, 4 vs 5, and 3 vs 5

eTable 1. Characteristics of the Development Set

eTable 2. Breakdown of Discordances (A) Between Subspecialists, (B) Between the Majority Opinion of Subspecialists and the Deep Learning System, DLS, and (C) Between the Majority Opinion of Subspecialists and General Pathologists

eTable 3. Individual Pathologist Gleason Scoring Agreement With Subspecialist Majority Opinion for the 19 Pathologists

eTable 4. DLS Performance For Gleason Grading Compared to the Majority Opinion of Subspecialists

eTable 5. Confusion Matrix (5 × 5 Contingency Table) Showing the Breakdown of Classifications for the DLS on Biopsies From ML2

eTable 6. Inter-pathologist Agreement on Tumor-Containing Slides

eTable 7. Sensitivity Analysis When Separating GG4 and GG5 into Separate Classification Categories

eTable 8. Hyperparameters for the Deep Learning System

eTable 9. Sensitivity of the DLS-Subspecialist Agreement in Tumor-Containing Cases to the Availability of Annotations From Each Individual Subspecialist

eTable 10. Sensitivity of the DLS-Subspecialist Agreement in Tumor-Containing Cases to the Availability of Annotations From Each Individual Subspecialist

eReferences

This supplementary material has been provided by the authors to give readers additional information about their work.

eMethods

Slide Preparation and Image Digitization

For each case in the validation set, fresh tissue sections were cut from deaccessioned tissue blocks beyond the 10-year Clinical Laboratory Improvement Amendments (CLIA) archival requirement. Five serial sections (of approximately 5-micron thickness) were cut in total from each block; sections 1, 3, and 5 were hematoxylin-and-eosin (H&E)-stained, while section 4 was triple-stained with the PIN4 immunohistochemistry cocktail. Slides from each of the 4 data sources (ML1, ML2, UH, and TTH) were cut and stained by 4 separate laboratories. In total, 1339 cases were initially scanned for the validation set; 752 were subsequently used based on urologic specialist review availability and exclusion criteria. Development set slides from ML1 followed a similar procedure to those above without obtaining a triple-stained PIN4 cocktail for each case, while development slides from TTH were obtained by scanning slides within the 10-year CLIA archival requirement. From UH, anonymized digital H&E slides were obtained. Slides from TTH, ML1, and ML2 were digitized for purposes of this study using a Leica Aperio AT2 scanner at a resolution of 0.25 $\mu\text{m}/\text{pixel}$ (“40X magnification”), while digital slides obtained from UH were each previously scanned on a Hamamatsu NanoZoomer S360 scanner at a resolution of 0.23 $\mu\text{m}/\text{pixel}$ (“40X magnification”) or 0.46 $\mu\text{m}/\text{pixel}$ (“20X magnification”).

Biopsy Reviews for DLS Development

For DLS development, 9 urologic subspecialist pathologists (A.E., A.S., C.C., J.S., M.A., M.Z., P.H., R.A., T.K.) assessed 524 biopsies, with a median of one review per biopsy (range 1-6). In addition to providing biopsy-level reviews as is performed in routine clinical practice, more precise glandular-level annotations were made to enable the DLS to recognize glandular level Gleason patterns. For these glandular annotations, board-certified general pathologists outlined individual glands or regions (eg, groups of glands) and annotated them as one of four categories: non-tumor, GP3, GP4, or GP5. These region-level categorizations were subsequently reviewed and corrected as appropriate by one of the nine subspecialists.

Glandular Annotations

Detailed “region-level annotations” that label glands or regions such as groups of glands were collected in a similar manner as previously described (see Supplemental Methods of previous study).¹ Annotations were performed in a custom histopathology viewer using free-drawing tools, typically between 5X and 20X magnifications (available range of magnification was 0.04X to 40X). Pathologists outlined regions as “Non-tumor”, and Gleason patterns (GP): “GP3”, “GP4”, and “GP5”. In cases of true histological ambiguity, annotators were given the ability to assign mixed-grades (e.g. “3+4”); these annotations were used at training time as the primary GP (e.g. “3”).

Deep Learning System

The Deep Learning System consists of two stages: a convolutional neural network (CNN) that classifies image patches within each biopsy, followed by a second machine learning model (a support vector machine, or SVM) that uses features extracted from the resulting heatmap to classify the biopsy’s overall Grade Group (GG). We first describe the development of the

custom CNN architecture for Gleason grading, followed by the training and tuning of the discovered network, and lastly the training and tuning of the second-stage SVM. Tensorflow² version 1.14.0 was used in construction of the convolutional neural network, while Scikit-learn³ version 0.20.0 was used for SVM development.

The first stage of the DLS operates on 128x128 μ m-sized regions referred to as image patches. Each image patch and its surrounding image context (total input image size, 512 μ m x 512 μ m) is evaluated by the deep convolutional neural network which provides its interpretation of the relative likelihood of each of four classes (non-tumor, GP3, GP4, and GP5) being present in the image patch.

For each slide, the region-level predictions are then assembled into a “heatmap” for the slide. The heatmap for the slide is then summarized by numerical features that characterize the biopsy: percentage of the prostate biopsy slide containing tumor and the relative percentages of each Gleason pattern. The second stage of the DLS uses these features as input to a second machine learning model (a support vector machine), to provide a GG classification for the biopsy.

Architecture Development

To develop a CNN architecture specifically for Gleason grading, we use a version of Neural Architecture Search (TuNAS).⁴ Briefly, the neural networks were defined by combining a set of modules, and each module had multiple different configurations. TuNAS programmatically searched through a prespecified configuration search space to create the final neural network architecture. The search space was constructed by specifying the number of modules in the network and allowing each module to vary among several predefined configurations. In each iteration, TuNAS sampled a neural network, evaluated the performance, and updated the parameters of the search algorithm. To estimate the performance of a sampled network, we trained the network and computed the loss function on a held-out subset of the development set. The final neural network used was obtained by selecting the configuration with the highest score for each module.

In the architecture search, a basis is required for the design of search space, termed a “backbone”. In this case, we used the Xception⁵ architecture, a performant network at image classification and segmentation tasks, and constructed a search space to allow for flexibility in the receptive field of the network.

Specifically, the Xception architecture consists of twelve total modules bracketed by skip connections⁶ (3 in the “entry flow”, 8 in the “middle flow” and 1 in the “exit flow”), with each module having two or three 3x3 convolutions. In the search space, we included alternate configurations in place of these ones: modules composed of 5x5 convolutions or 7x7 convolutions. Similarly, the search space also included the choice of swapping the last two 3x3 convolutions for two 5x5 convolutions or two 7x7 convolutions respectively. Skipping of the “middle flow” modules (i.e. an identity operation module) was also permitted such that the search could trade off depth and width as necessary. As such, the search space consisted of approximately 16 million possible architectures, one of which is the original Xception network.

The architecture search was conducted using the dataset (from previous work) for Gleason Grading of prostatectomies because of the larger number of glandular (“region-level”) annotations in that dataset.¹ This dataset was split into training and tuning sets as previously described: 3 million patches were sampled from the training set for use as the search process’s

training set, and 1 million patches were subsampled from the tuning set for use as the search process's tuning set. Hyperparameters for the search are presented in eTable 8, and the discovered network is presented in eFigure 1.

Architecture Training and Ensembling

The top discovered architecture was then retrained and tuned using the full prostatectomy development and validation sets from the previous study. As previously described¹, color augmentations, orientation randomization, and stain normalization were employed to improve performance, and hyperparameters were tuned using Google Vizier⁷. See eTable 8 for resulting hyperparameters.

Next, the network was refined using annotated biopsies (eTable 1 below). Annotated biopsy slides were randomly split into three folds, and three separate networks were initialized from the same prostatectomy-trained weights and refined using each of the dataset folds. In addition to color augmentation, orientation randomization, and stain normalization, cutout augmentations⁸ were additionally used to improve model performance. Hyperparameters for each fold were tuned using Google Vizier (eTable 8).⁷ An ordinal loss function⁹ was used for training and refinement.

Finally, at evaluation time, nine models were trained and ensembled (three models for each of the three folds) by taking the geometric mean across all model predictions for each patch.

Thresholding and Stage 2 Features

The DLS's first stage assigned the probabilities (in the range [0, 1]) of each patch to be one of four classes: non-tumor or GP3, GP4, or GP5. To map these probabilities to a predicted class, we thresholded the predictions. First, a patch was categorized as non-tumor if the predicted non-tumor probability exceeded 0.2. Otherwise, the top two GPs' predicted probabilities were re-normalized to sum to 1.0, and compared against a threshold based on the specific GPs. The thresholds were 0.65 for GP3/4, 0.94 for GP 3/5, and 0.90 for GP4/5; the more severe GP was assigned if the threshold was exceeded. These thresholds were selected empirically via 10-fold cross validation on the development set to optimize slide-level agreement with subspecialist-provided Gleason pattern percentages.

Features were then extracted from both the predicted probabilities for each patch and the 4-class categorization. A SVM then used these features to classify each biopsy as: non-tumor, GG1, GG2, GG3, or GG4-5. The features were the percent of biopsy classified as non-tumor, percent of tumor classified as GP4, and GP5 respectively, the lowest predicted patch-wise non-tumor probability, and the 98th percentile of the patch-wise predicted probabilities for GP4 and GP5 respectively. Hyperparameters for the SVM were tuned using 10-fold cross validation across the biopsy-level dataset (eTable1) and are presented in eTable 8. The predicted probabilities of the SVM for each category were summed for the purposes of receiver operating characteristic (ROC) analyses. For example, among non-tumor cases, plotting the ROC of GG1-2 vs GG3-5 involved summing for each case the SVM's predicted probability values of GG1 and GG2, versus GG3 and GG4-5.

Statistical Analysis

To evaluate the DLS, it was compared to the majority-vote of at least 2 expert subspecialists from a panel of expert 6 (Methods section of main text). We used this majority-vote approach instead of panel review based on preliminary data for 50 biopsies (independent of the validation dataset) suggesting that panel discussion rarely helped resolve disagreements

between experts, as well as being extremely time-consuming and therefore impractical for a dataset of this size.

To compute 95% confidence intervals, we used a slide resampling bootstrap approach. In each iteration of the bootstrap, we sampled with replacement a set of slides of the same size as the original set, and compute the metric of interest. After 1000 iterations, we report the 2.5th and 97.5th percentiles as the confidence interval bounds.

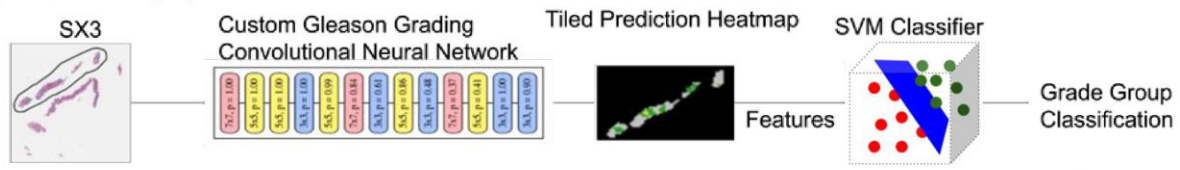
With regard to Grade Group Classification, we performed several sub-analyses and sensitivity analyses. First, to evaluate generalization of the DLS to a datasource with different staining, scanning, and patient characteristics, we considered biopsies from the external validation set only (ML2, n=175, Main Figure 1). Next, we considered the subset of tumor-containing cases with GG consensus among subspecialists (n=328, eTable 4). Then, we further carried out an analysis treating GG4 and GG5 separately, rather than combining them into a single category (eTable 7). Finally, sensitivity analysis of the DLS-subspecialist agreement with respect to individual subspecialists' annotations are provided in eTables 9 and 10, and sensitivity analysis of the DLS agreement with respect to the DLS-internal Gleason pattern thresholding is presented in eFigure 6.

The DLS's Gleason grading agreement with the majority opinion of subspecialists was additionally evaluated by area under the receiver operating characteristic curve (Area under ROC, AUC) analysis. The AUCs were estimated using the Wilcoxon (Mann-Whitney) U statistic, a standard nonparametric method employed by most modern software libraries. To obtain binary outcomes necessary for AUC analysis, the five categories of Gleason scores were dichotomized using clinically important cutoffs. Specifically, we used ROC analysis to evaluate DLS grading of slides as GG1 vs. GG2-5, a distinction representing the clinically significant threshold for potential eligibility for active surveillance versus prostatectomy/definitive treatment^{10,11}. We also evaluated the tumor versus non-tumor threshold to represent the important diagnostic step of establishing a prostatic adenocarcinoma diagnosis. Lastly we evaluated GG1-2 versus GG3-5 as some patients with GG2 may still be managed with active surveillance if only a very low amount of Gleason pattern 4 was present^{10,11}.

(A) Urologic Subspecialists



(B) Deep Learning System

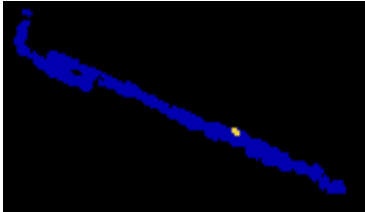
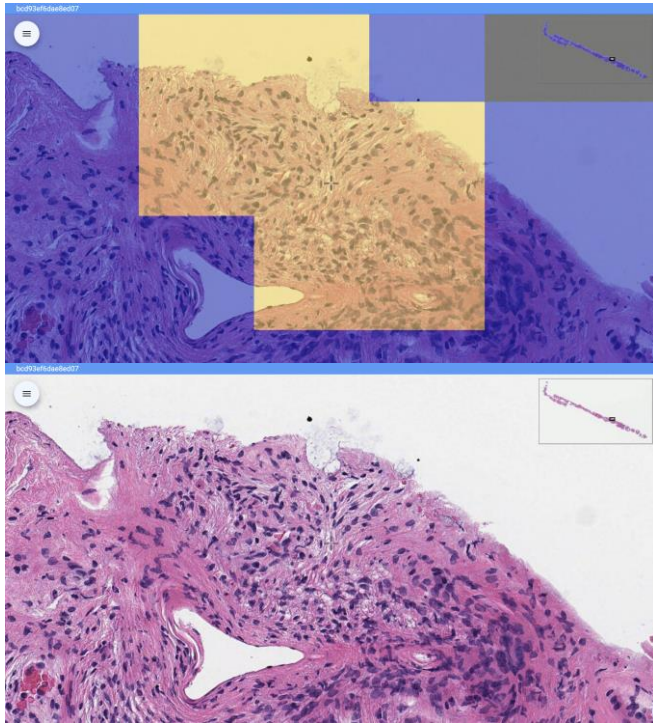


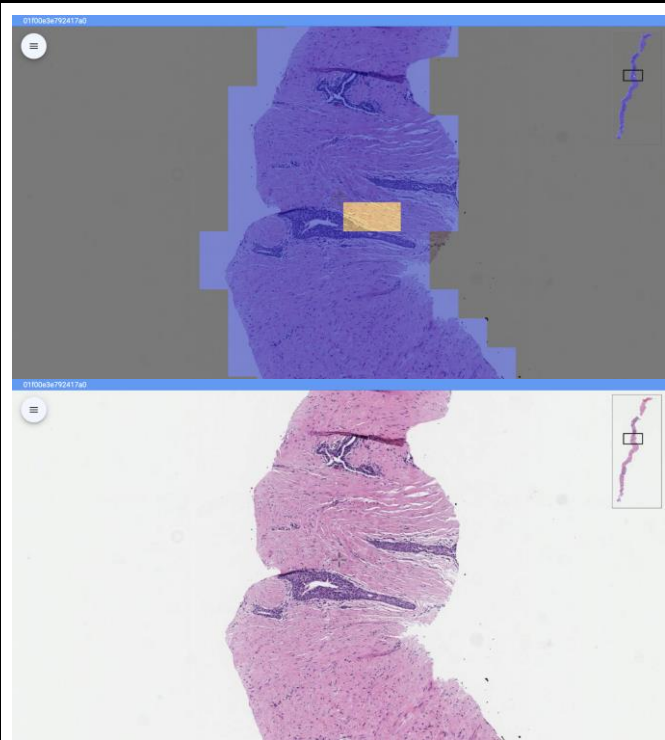
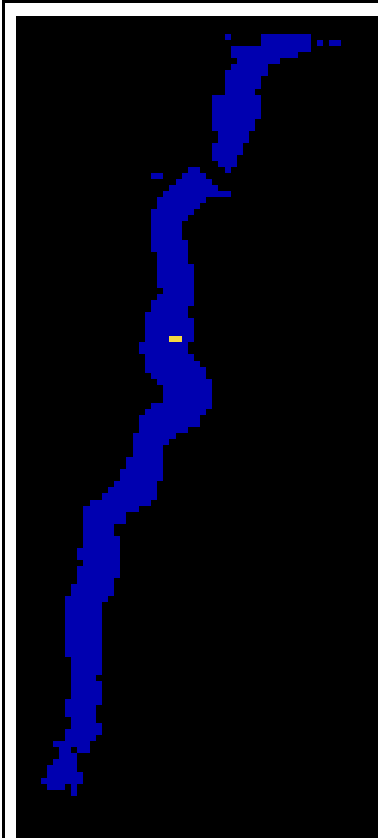
(C) Cohort of General Pathologists



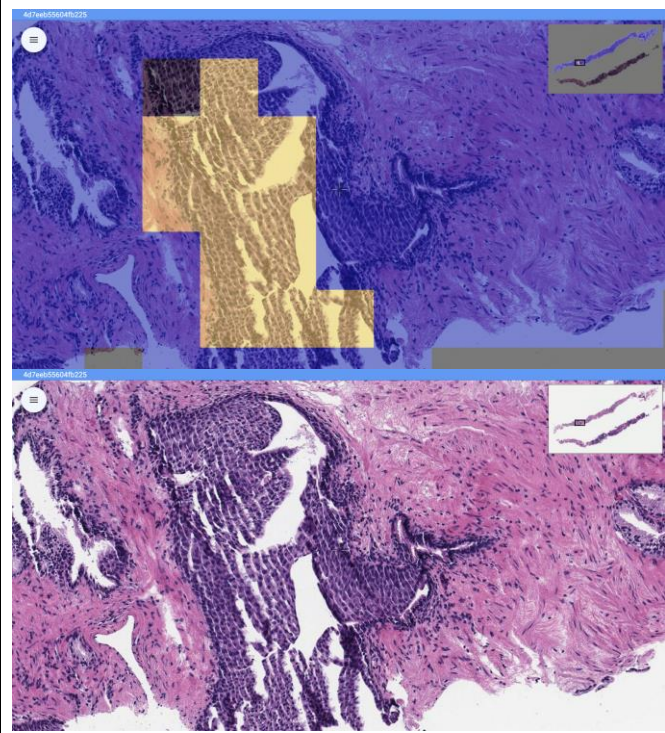
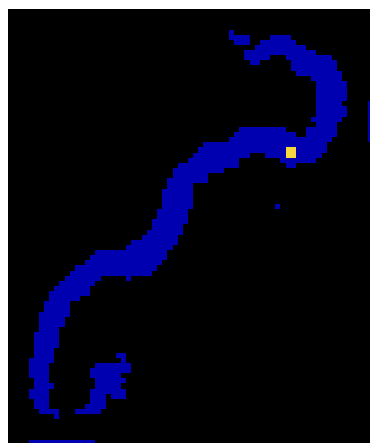
eFigure 1. Overview of the reviews provided by expert subspecialists for the validation set and an overview of the deep learning system.

(A) Six urologic expert subspecialists with an average of 25 years of experience (range: 18-34) contributed to the subspecialists' majority opinion in this study. Levels and immunohistochemistry were available for every biopsy to reduce diagnostic uncertainty around "tangential cuts" and non-tumor vs. tumor delineations. Two specialists initially reviewed each slide, and in cases of classification disagreements (on non-tumor vs. Grade Groups 1, 2, 3, or 4-5), a third specialist review was collected. **(B)** Overview of the two-stage deep learning system and customized neural network architecture for Gleason grading. The search space consisted of thirteen modules, each with several possible configurations, resulting in approximately 16 million possible architectures in total (see "Architecture Development"). Each of the discovered architecture's finalized thirteen configurations is highlighted in pink, yellow, and blue for modules containing 7x7, 5x5, and 3x3 convolutions respectively. **(C)** A cohort of 19 pathologists provided reviews on overlapping subsets of the validation set with access to multiple levels for each biopsy to reflect typical clinical workflows.

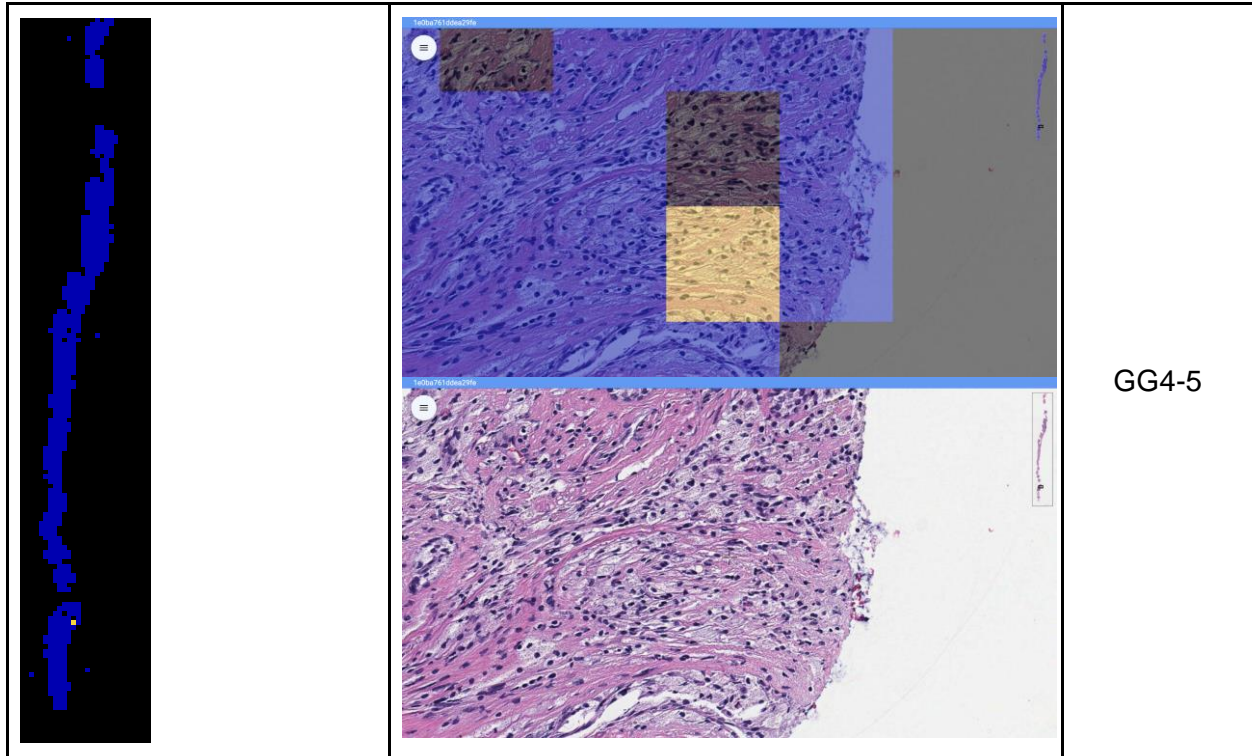
<p>DLS predicted glandular Gleason pattern distribution</p> <ul style="list-style-type: none"> ■ Non-tumor ■ GP 3 ■ GP 4 ■ GP 5 	<p>Example tumor area</p>	<p>DLS- predicted GG for the biopsy</p>
		<p>GG3</p>



GG3

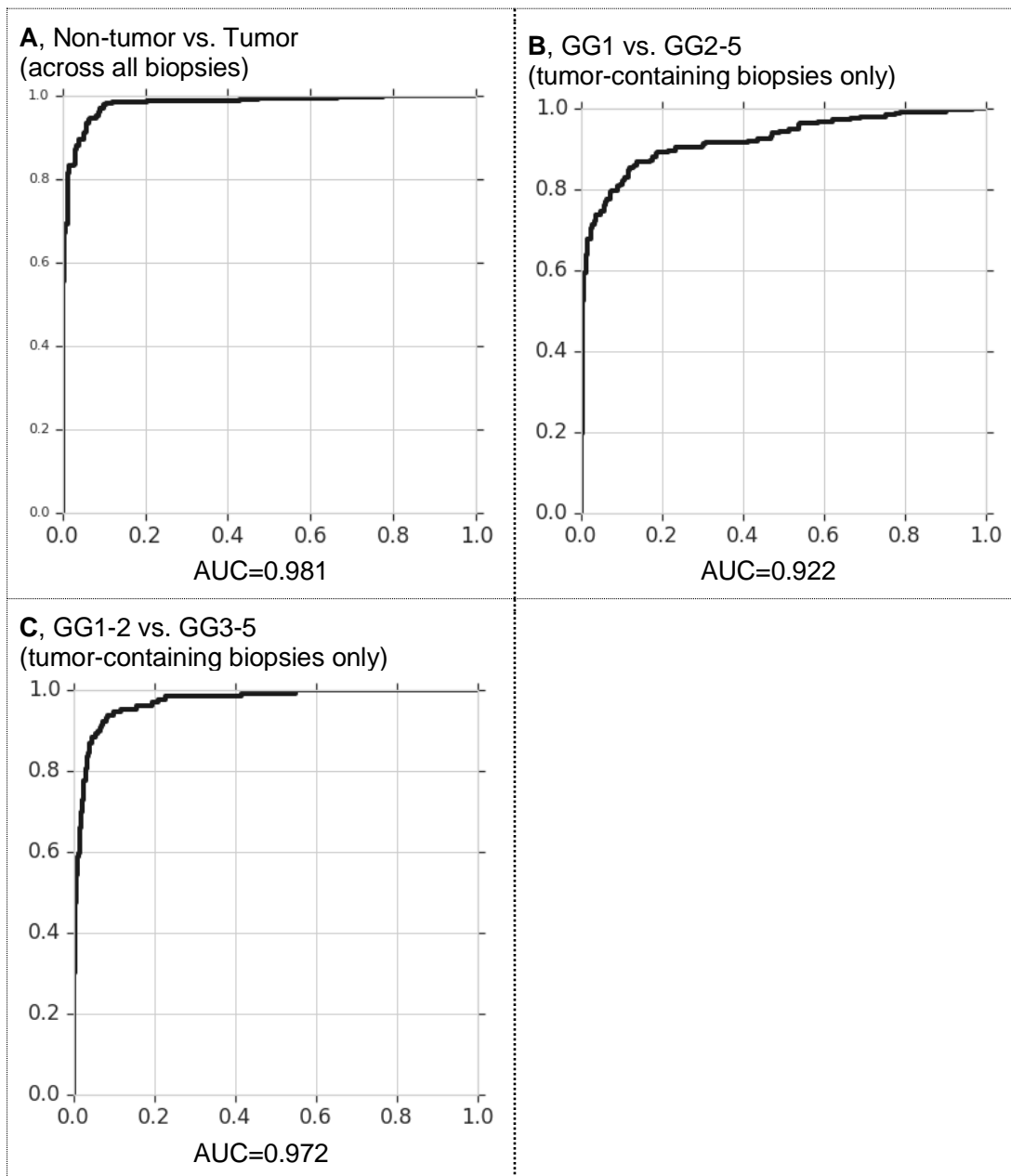


GG4-5



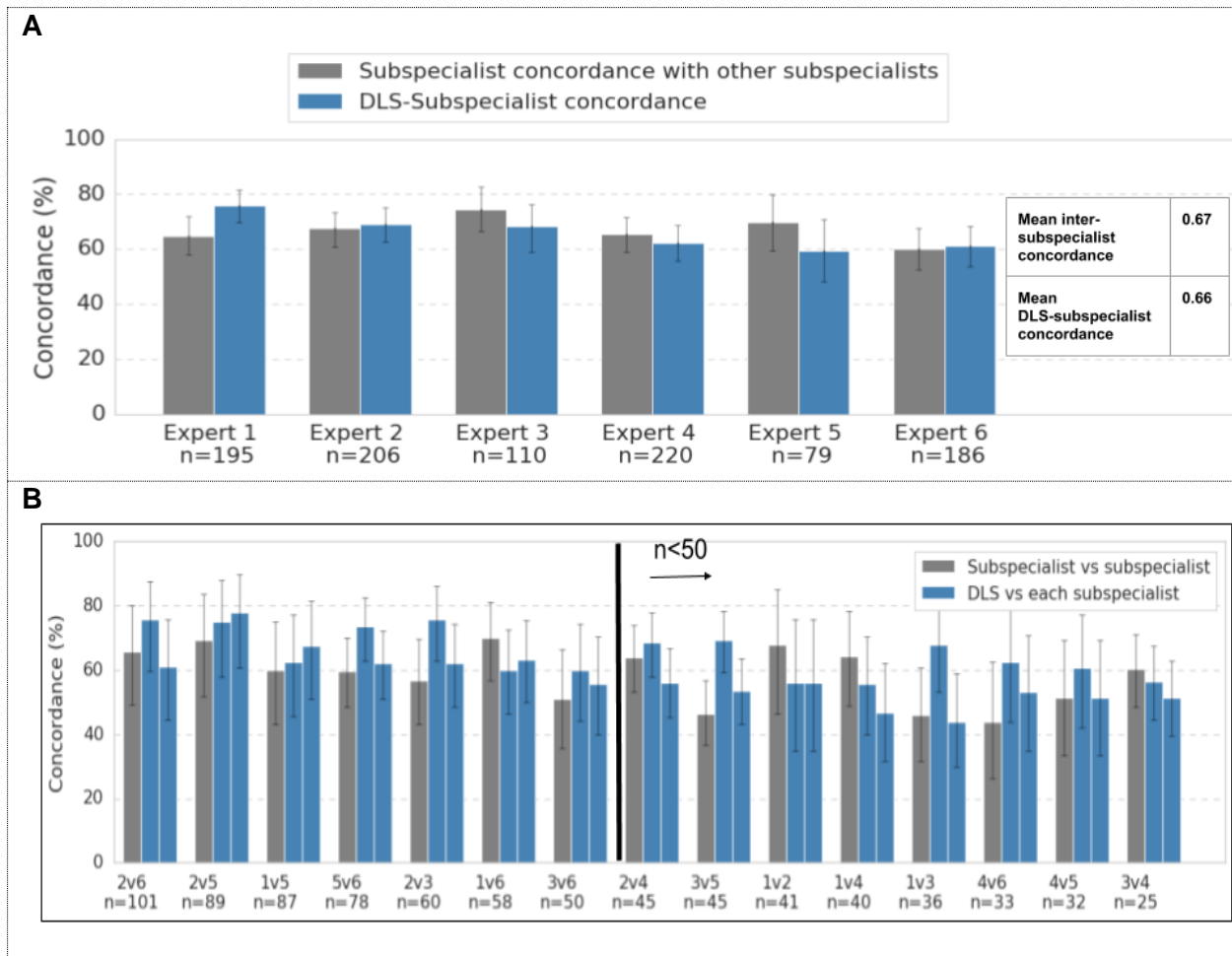
eFigure 2. Example Overcalls by the DLS on Biopsies Without Tumor

Overcalls were generally on small regions near tissue artifacts. These overcalls are unlikely to mislead pathologists using the DLS as a decision support tool by also reviewing these gland-level predictions.



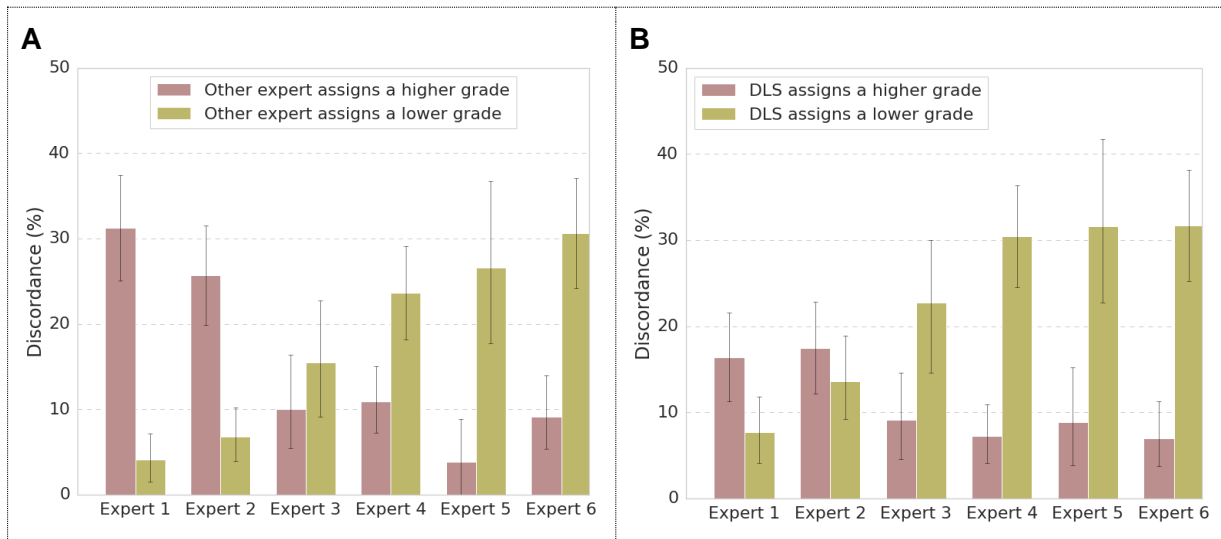
eFigure 3. Validation Set Performance for Gleason Grading and Tumor Detection

Each plot shows the receiver operating characteristic curves of sensitivity and specificity of the DLS. The DLS made five-category determinations for each case. To obtain the sensitivity and specificity, we grouped the Grade Group determinations for binary analysis. For example, a categorization of GG4 would be considered a true positive for the grouping GG3-5 versus GG1-2. The DLS was able to provide a predicted probability for each of the five categories, and the predicted probability for each grouping was the sum of all predicted probabilities within the group, such as GG1 and GG2 for the group GG1-2. This grouped predicted probability was used to plot the continuous receiver operating characteristics curves. Three clinically important cutoffs are shown: non-tumor versus tumor (A); GG1 versus GG2-5 on tumor-containing slides (B); and GG1-2 versus GG3-5 on tumor-containing slides (C).



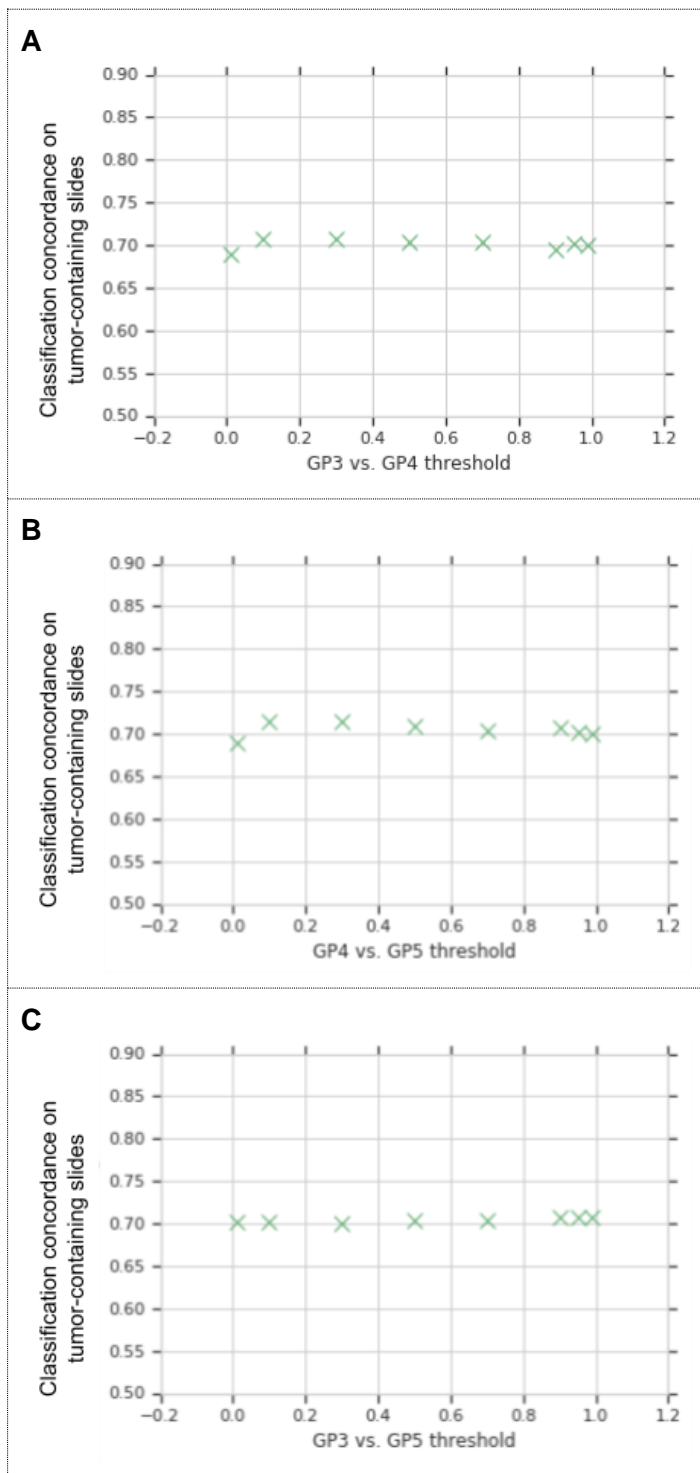
eFigure 4. Evaluation of Inter-Subspecialist Agreements

(A) Agreement of each individual subspecialist with the other subspecialist that graded the same case. (B) Head-to-head comparison of subspecialist:subspecialist agreement with DLS:subspecialist agreement on tumor-containing slides. Each subspecialist graded a subset of the images, with every image being graded by two of the six subspecialists. The grey bars show all 15 (6 choose 2) pairwise inter-subspecialist agreement measurements on the cases graded in common. The two blue bars paired with each grey bar represent the agreement of the DLS with each of the associated two subspecialists, on the exact same set of images. The pairs of subspecialists are sorted in order of decreasing number of images, with $n < 50$ on the right. The mean inter-subspecialist agreement (66%) is greater than the mean inter-pathologist agreement (53%) provided in eTable 6 below.



eFigure 5. Comparison of the Deep Learning System (DLS) With Individual Subspecialists on Tumor-Containing Slides

(A) Break down of discordances between individual expert subspecialists. Relative to the individual subspecialist, the blue and red bars represent the proportion of cases where the other subspecialist assigned a higher and lower grade, respectively. (B) Break down of DLS:subspecialist discordances. Relative to the individual subspecialist, the blue and red bars represent the proportion of cases where the DLS assigned a higher and lower grade, respectively. Error bars represent the 95% confidence intervals. Potential reasons for disagreements among expert subspecialists may include inherent uncertainty in the two-dimensional interpretation of a three-dimensional specimen, ambiguity in grading guidelines, inexactness of visual quantitation, and cognitive factors such as anchoring. However, this analysis suggests that some discordances may be due to systematic differences in “grading calibration” (consistently over- or under-grading relative to other subspecialists).



eFigure 6. Sensitivity of the DLS’s Gleason Grade Group Classification Accuracy to Differences in Gleason Pattern (GP) Thresholds for GP 3 vs 4, 4 vs 5, and 3 vs 5

See “Thresholding and Stage 2 Features” in the Supplementary Methods for more details.

eTable 1. Characteristics of the Development Set

The development set contains prostate biopsy cases from a large tertiary teaching hospital (TTH), a medical laboratory (ML1), and a University Hospital (UH). Biopsy-level pathologic reviews were obtained from ML1 and TTH, while detailed region-level annotations were obtained from all three sources.

Biopsy-Level Reviews				
Urologic subspecialist reviews	Medical Laboratory 1	Tertiary Teaching Hospital		Total
Non-tumor	72	50		122
Grade Group 1	30	172		202
Grade Group 2	19	111		120
Grade Group 3	5	42		47
Grade Group 4-5	37	42		79
Total	165 reviews / 135 biopsies / 135 cases	417 reviews / 389 biopsies / 225 cases		580 reviews / 524 biopsies / 360 cases
Region-Level Annotated Biopsy Patches				
Urologic subspecialist reviews	Medical Laboratory 1	Tertiary Teaching Hospital	University Hospital	Total
Non-tumor	182,938	620,916	495,715	1,299,569
Gleason Pattern 3	15,790	43,998	82,740	142,528
Gleason Pattern 4	28,207	112,120	59,897	200,224
Gleason Pattern 5	2,742	28,158	8,066	38,966
Total	229,677 patches / 73 biopsies	805,192 patches / 156 biopsies	646,418 patches / 115 biopsies	1,681,287 patches / 344 biopsies

eTable 2. Breakdown of Discordances (A) Between Subspecialists, (B) Between the Majority Opinion of Subspecialists and the Deep Learning System, DLS, and (C) Between the Majority Opinion of Subspecialists and General Pathologists

The diagonal values (in bold) represent agreement; the region below the diagonal represents relative undergrading; and the region above the diagonal right region represents relative overgrading. Relative to the majority opinion of subspecialists, the Cohen's kappa for the DLS was 0.71, while the mean Cohen's kappa amongst general pathologist was 0.61.

A

Subspecialist Classification	Subspecialist Classification				
	Non-tumor	GG1	GG2	GG3	GG4-5
Non-tumor	33.8%	1.9%	0.0%	0.0%	0.0%
GG1		25.0%	10.6%	0.9%	0.1%
GG2			8.2%	4.8%	0.2%
GG3				4.3%	3.9%
GG4-5					6.3%

B

Majority opinion of subspecialists	DLS (n=752 reviews across 752 biopsies)				
	Non-tumor	GG1	GG2	GG3	GG4-5
Non-tumor	31.0%	0.9%	0.4%	1.5%	0.0%
GG1	2.7%	26.2%	3.7%	0.3%	0.0%
GG2	0.0%	4.7%	10.1%	1.5%	0.0%
GG3	0.3%	0.3%	2.4%	6.0%	0.4%
GG4-5	0.0%	0.0%	0.1%	2.5%	5.2%

C

Majority opinion of subspecialists	Pathologists (n=2239 reviews across 752 biopsies)				
	Non-tumor	GG1	GG2	GG3	GG4-5
Non-tumor	32.5%	0.8%	0.0%	0.0%	0.0%
GG1	3.4%	20.4%	7.1%	0.9%	0.4%
GG2	0.2%	3.5%	8.1%	3.7%	1.2%
GG3	0.2%	0.2%	1.3%	4.1%	3.8%
GG4-5	0.1%	0.0%	0.1%	0.8%	7.1%

eTable 3. Individual Pathologist Gleason Scoring Agreement With Subspecialist Majority Opinion for the 19 Pathologists

Each pathologist reviewed overlapping subsets of the validation set. Agreement with the majority opinion in classifying each slide (as non-tumor, or Grade Groups 1, 2, 3, or 4-5) ranged from 49-87% (mean: 72%). On tumor-containing biopsies, accuracies ranged from 27-75% (mean: 58%). Bold indicates the higher accuracy within all slides or tumor-containing slides.

Pathologist	Among all slides			Among tumor-containing slides only		
	Agreement with subspecialist majority opinion (% 95% CIs)	DLS agreement with subspecialist majority opinion on same subset of slides (95% CIs)	Number of slides	Agreement with subspecialist majority opinion (% 95% CIs)	DLS agreement with subspecialist majority opinion on same subset of slides (95% CIs)	Number of slides
1	48.8 (34.9, 62.8)	74.4 (60.5, 86.0)	43	26.7 (10.0, 43.3)	66.7 (50.0, 83.3)	30
2	59.5 (48.6, 71.6)	71.6 (60.8, 81.1)	74	49.1 (36.4, 61.8)	61.8 (49.1, 74.5)	55
3	61.6 (51.2, 72.1)	79.1 (69.8, 87.2)	86	55.9 (44.1, 67.6)	77.9 (67.6, 88.2)	68
4	64.7 (59.3, 69.9)	73.4 (68.6, 78.2)	312	62.1 (56.2, 68.0)	71.7 (66.2, 76.8)	272
5	66.7 (57.6, 75.8)	83.8 (75.8, 90.9)	99	43.6 (30.9, 58.2)	74.5 (63.6, 85.5)	55
6	69.2 (62.5, 76.0)	79.8 (74.0, 85.6)	208	51.5 (43.1, 60.0)	71.5 (63.8, 79.2)	130
7	70.2 (55.3, 83.0)	80.9 (68.1, 91.5)	47	63.2 (47.4, 78.9)	76.3 (63.2, 89.5)	38
8	71.4 (61.9, 81.0)	78.6 (69.0, 86.9)	84	50.0 (35.4, 64.6)	70.8 (56.2, 83.3)	48
9	71.9 (59.4, 82.8)	82.8 (73.4, 90.6)	64	61.7 (48.9, 74.5)	76.6 (63.8, 89.4)	47
10	72.3 (64.8, 79.2)	74.8 (67.9, 81.8)	159	61.7 (53.0, 70.4)	69.6 (61.7, 77.4)	115
11	72.7 (66.7, 77.9)	76.3 (71.1, 81.1)	249	61.1 (53.1, 67.9)	67.9 (61.1, 74.7)	162
12	74.1 (64.2, 82.7)	77.8 (67.9, 86.4)	81	64.4 (52.5, 76.3)	76.3 (66.1, 86.4)	59
13	75.6 (61.0, 87.8)	78.0 (63.4, 90.2)	41	63.0 (44.4, 81.5)	70.4 (51.9, 85.2)	27
14	76.1 (67.3, 84.1)	81.4 (74.3, 87.6)	113	48.0 (34.0, 62.0)	66.0 (52.0, 80.0)	50
15	77.9 (67.6, 86.8)	79.4 (69.1, 88.2)	68	58.3 (41.7, 75.0)	63.9 (47.2, 80.6)	36
16	81.6 (73.5, 88.8)	77.6 (69.4, 85.7)	98	75.0 (65.3, 84.7)	76.4 (66.7, 86.1)	72
17	81.7 (77.5, 85.6)	81.4 (76.8, 85.6)	306	67.1 (60.0, 73.5)	72.9 (66.5, 79.4)	170
18	83.6 (73.8, 91.8)	86.9 (77.0, 95.1)	61	74.4 (59.0, 87.2)	82.1 (69.2, 92.3)	39
19	87.0 (76.1, 95.7)	80.4 (67.4, 91.3)	46	64.7 (41.2, 88.2)	47.1 (23.5, 70.6)	17

eTable 4. DLS Performance For Gleason Grading Compared to the Majority Opinion of Subspecialists

Numbers indicate agreement with the majority opinion, with 95% confidence intervals in parenthesis, and sample size reported.

	Entire validation set	Subset where first 2 subspecialists were concordant	Subset from independent data source (ML2)
All biopsies	0.785 (0.755, 0.814); n=752	0.835 (0.804, 0.864); n=576	0.801 (0.757-0.845); n=322
Tumor-containing biopsies	0.717 (0.679, 0.753); n=503	0.774 (0.729, 0.817); n=328	0.714 (0.657, 0.777); n=175

eTable 5. Confusion Matrix (5 × 5 Contingency Table) Showing the Breakdown of Classifications for the DLS on Biopsies From ML2

Majority opinion of subspecialists	DLS (n=322 reviews across 322 biopsies)				
	Non-tumor	GG1	GG2	GG3	GG4-5
Non-tumor	41.3%	0.9%	0.9%	2.5%	0.0%
GG1	2.2%	17.1%	3.7%	0.62%	0.0%
GG2	0.0%	3.1%	8.4%	2.2%	0.0%
GG3	0.6%	0.0%	0.6%	5.3%	0.3%
GG4-5	0.0%	0.0%	0.0%	2.2%	8.1%

eTable 6. Interpathologist Agreement on Tumor-Containing Slides

Inter-pathologist agreement, grouped by every individual pathologist. Each individual pathologist's reviews are compared against all reviews from other pathologists available on the same cases. Resulting inter-pathologist agreement ranges from 38-72% with a mean inter-pathologist agreement of 52%. The mean inter-pathologist agreement (53%) is lower than the mean inter-subspecialist agreement (66%) provided in eFigure 4 above.

Pathologist	Agreement with other pathologists (95% CIs)	Number of comparisons
1	38.6 (27.1, 50.0)	70
2	37.7 (30.0, 46.9)	130
3	43.5 (35.1, 51.9)	131
4	49.2 (45.0, 53.6)	524
5	45.8 (37.5, 54.2)	120
6	54.5 (48.8, 60.7)	242
7	50.0 (39.3, 59.5)	84
8	50.0 (39.6, 58.5)	106
9	51.9 (42.6, 61.1)	108
10	60.7 (54.1, 66.9)	242
11	53.7 (48.7, 58.7)	339
12	50.0 (41.4, 58.6)	128
13	55.6 (44.4, 66.7)	63
14	58.2 (49.1, 67.3)	110
15	57.8 (48.0, 66.7)	102
16	64.5 (57.2, 71.7)	152
17	55.0 (49.6, 59.8)	353
18	72.3 (62.7, 80.7)	83
19	48.8 (34.1, 63.4)	41

eTable 7. Sensitivity Analysis When Separating GG4 and GG5 into Separate Classification Categories

In the main analyses, GG4-5 were combined due to their low incidence and often similar treatment implications. A sensitivity analysis treating GG4 and GG5 separately is presented here. In this new analysis, 11 of the 752 biopsies were excluded because the first two subspecialists' reviews were discordant between GG4 and GG5 and so a third review wasn't collected.

A

Subspecialists' majority opinion	Deep Learning System					
	Non-tumor	GG1	GG2	GG3	GG4	GG5
Non-tumor	32.1%	0.8%	0.3%	0.5%	0.3%	0.3%
GG1	3.2%	26.5%	3.4%	0.1%	0.1%	0.0%
GG2	0.4%	5.0%	9.6%	1.3%	0.1%	0.0%
GG3	0.3%	0.4%	2.4%	5.4%	0.4%	0.5%
GG4	0.0%	0.0%	0.0%	0.8%	1.1%	0.9%
GG5	0.0%	0.0%	0.0%	0.3%	0.3%	3.1%

B

Subspecialists' majority opinion	General pathologists					
	Non-tumor	GG1	GG2	GG3	GG4	GG5
Non-tumor	33.1%	0.9%	0.0%	0.0%	0.0%	0.0%
GG1	3.5%	20.7%	7.2%	0.9%	0.4%	0.1%
GG2	0.2%	3.6%	8.2%	3.7%	1.0%	0.3%
GG3	0.2%	0.2%	1.3%	4.1%	2.6%	1.2%
GG4	0.1%	0.0%	0.0%	0.3%	1.8%	0.7%
GG5	0.0%	0.0%	0.0%	0.3%	0.7%	2.6%

eTable 8. Hyperparameters for the Deep Learning System

Headings are bolded for visual clarity.

Architecture search hyperparameters			
Neural network learning rate schedule	Cosine decay with linear warmup schedule Base rate: 4.2×10^{-3} Decay steps: 50000 Fraction of training steps used for linear warmup: 0.025		
Neural network RMSProp optimizer	Decay: 0.9 Momentum: 0.9 Epsilon: 1.0		
Controller Adam optimizer	Base rate: 2.5×10^{-4} Momentum: 0.95 Beta1: 0.000 Beta2: 0.999 Epsilon: 1×10^{-8}		
Batch size	128		
Network pre-training hyperparameters (prostatectomy data)			
Color perturbations	Saturation delta: 0.80 Brightness delta: 0.96 Contrast delta: 0.17 Hue delta: 0.02		
Learning rate schedule	Exponential decay schedule Base rate: 0.0042 Decay rate: 0.95 Decay steps: 51,733 steps		
RMSProp optimizer	Decay: 0.95 Momentum: 0.7 Epsilon: 0.001		
Other	Loss function: softmax cross-entropy Batch size: 32		
Network refinement hyperparameters (biopsy data)			
	Fold 1	Fold 2	Fold 3

Image augmentations	Saturation delta: 0.53 Brightness delta: 0.32 Contrast delta: 0.61 Hue delta: 0.01 Cutout box size: 50x50 pixels		
Learning rate schedule (exponential decay schedule)	Base rate: 2.3×10^{-5} Decay rate: 0.70 Decay steps: 72,466	Base rate: 3.2×10^{-5} Decay rate: 0.50 Decay steps: 75,936	Base rate: 3.8×10^{-5} Decay rate: 0.95 Decay steps: 28,512
RMSProp optimizer	Decay: 0.90 Momentum: 0.90 Epsilon: 1.00	Decay: 0.95 Momentum: 0.90 Epsilon: 1.0	Decay: 0.95 Momentum: 0.70 Epsilon: 0.10
Other	Loss function: Ordinal cross-entropy Batch size: 32		
Support Vector Machine hyperparameters			
Penalty parameter ('C')	100		
Kernel	RBF, Gamma = 0.25		

eTable 9. Sensitivity of the DLS-Subspecialist Agreement in Tumor-Containing Cases to the Availability of Annotations From Each Individual Subspecialist

Subspecialist	Original DLS agreement with subspecialist (95% CI)	DLS agreement with subspecialist if trained without annotations from this subspecialist (95% CI)	Number of validation cases reviewed by the subspecialist	Number of excluded training cases reviewed by the subspecialist (%)
Subspecialist 1	0.721 (0.655, 0.782)	0.697 (0.635, 0.764)	n=195	n=162 (21.2%)
Subspecialist 2	0.673 (0.606, 0.740)	0.646 (0.578, 0.709)	n=206	n=105 (13.7%)
Subspecialist 3	0.658 (0.568, 0.748)	0.654 (0.545, 0.672)	n=110	n=30 (3.93%)
Subspecialist 4	0.621 (0.558, 0.683)	0.609 (0.545, 0.673)	n=220	n=55 (7.22%)
Subspecialist 5	0.613 (0.500, 0.725)	0.632 (0.518, 0.734)	n=79	n=58 (7.61%)
Subspecialist 6	0.597 (0.527, 0.672)	0.521 (0.451, 0.602)	n=186	n=22 (2.89%)

eTable 10. Sensitivity of the DLS-Subspecialist Agreement in Tumor-Containing Cases to the Availability of Annotations From Each Individual Subspecialist

Subspecialist	DLS agreement with the subspecialists' majority opinion (95% CI)
Including all subspecialists	0.717 (0.679, 0.753)
Excluding subspecialist 1	0.680 (0.636, 0.720)
Excluding subspecialist 2	0.668 (0.626, 0.708)
Excluding subspecialist 3	0.690 (0.646, 0.730)
Excluding subspecialist 4	0.674 (0.630, 0.712)
Excluding subspecialist 5	0.694 (0.652, 0.734)
Excluding subspecialist 6	0.692 (0.650, 0.732)

eReferences

1. Nagpal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* **2**, 48 (2019).
2. Abadi, M. et al. TensorFlow: A system for large-scale machine learning. (2016).
3. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
4. Bender, G. et al. Can weight sharing outperform random architecture search? An investigation with TuNAS. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). In press.
5. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017) doi:10.1109/cvpr.2017.195.
6. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) doi:10.1109/cvpr.2016.90.
7. Golovin, D. et al. Google Vizier. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17* (2017) doi:10.1145/3097983.3098043.
8. DeVries, T. & Taylor, G. W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv [cs.CV]* (2017).
9. Frank, E. & Hall, M. A Simple Approach to Ordinal Classification.
10. Chen, R. C., Bryan Rumble, R. & Jain, S. Active Surveillance for the Management of Localized Prostate Cancer (Cancer Care Ontario guideline): American Society of Clinical Oncology Clinical Practice Guideline Endorsement Summary. *Journal of Oncology Practice* vol. 12 267–269 (2016).
11. Morash, C. et al. Active surveillance for the management of localized prostate cancer: Guideline recommendations. *Can. Urol. Assoc. J.* **9**, 171–178 (2015).