Journal of Pathology Informatics 15 (2024) 100381

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

A selective CutMix approach improves generalizability of deep learning-based grading and risk assessment of prostate cancer

Sushant Patkar [a,1], Stephanie Harmon [a,*,1], Isabell Sesterhenn [b], Rosina Lis [a], Maria Merino [c], Denise Young [d,f], G. Thomas Brown [a], Kimberly M. Greenfield [b], John D. McGeeney [b], Sally Elsamanoudi [d,f], Shyh-Han Tan [d,f], Cara Schafer [d,f], Jiji Jiang [d,f], Gyorgy Petrovics [d,f], Albert Dobi [d,f], Francisco J. Rentas [b], Peter A. Pinto [g], Gregory T. Chesnut [d,e,h], Peter Choyke [a], Baris Turkbey [a], Joel T. Moncur [b]

[a] Artificial Intelligence Resource, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
[b] The Joint Pathology Center, Silver Spring, MD 20910, USA
[c] Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[d] Center for Prostate Disease Research, Murtha Cancer Center Research Program, Department of Surgery, Uniformed Services University of the Health Sciences, Bethesda, MD 20817, USA
[e] F. Edward Hebert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA
[f] Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD 20817, USA
[g] Urologic Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
[h] Urology Service, Walter Reed National Military Medical Center, Bethesda, MD 20814, USA

ARTICLE INFO

Keywords:
Prostate cancer
Digital pathology
Gleason grading
Artificial intelligence

ABSTRACT

The Gleason score is an important predictor of prognosis in prostate cancer. However, its subjective nature can result in over- or under-grading. Our objective was to train an artificial intelligence (AI)-based algorithm to grade prostate cancer in specimens from patients who underwent radical prostatectomy (RP) and to assess the correlation of AI-estimated proportions of different Gleason patterns with biochemical recurrence-free survival (RFS), metastasis-free survival (MFS), and overall survival (OS). Training and validation of algorithms for cancer detection and grading were completed with three large datasets containing a total of 580 whole-mount prostate slides from 191 RP patients at two centers and 6218 annotated needle biopsy slides from the publicly available Prostate Cancer Grading Assessment dataset. A cancer detection model was trained using MobileNetV3 on 0.5 mm × 0.5 mm cancer areas (tiles) captured at 10× magnification. For cancer grading, a Gleason pattern detector was trained on tiles using a ResNet50 convolutional neural network and a selective CutMix training strategy involving a mixture of real and artificial examples. This strategy resulted in improved model generalizability in the test set compared with three different control experiments when evaluated on both needle biopsy slides and whole-mount prostate slides from different centers. In an additional test cohort of RP patients who were clinically followed over 30 years, quantitative Gleason pattern AI estimates achieved concordance indexes of 0.69, 0.72, and 0.64 for predicting RFS, MFS, and OS times, outperforming the control experiments and International Society of Urological Pathology system (ISUP) grading by pathologists. Finally, unsupervised clustering of test RP patient specimens into low-, medium-, and high-risk groups based on AI-estimated proportions of each Gleason pattern resulted in significantly improved RFS and MFS stratification compared with ISUP grading. In summary, deep learning-based quantitative Gleason scoring using a selective CutMix training strategy may improve prognostication after prostate cancer surgery.

Introduction

Prostate cancer is characterized by a broad spectrum of clinical behavior spanning indolent to highly aggressive disease. Clinical risk assessment in localized prostate cancer largely depends on pathological assessment of morphological appearance with the Gleason scoring system.[1] However, prostate cancer exhibits both intra- and inter-tumor heterogeneity of growth patterns, with increasing morphological diversity in larger cancers of worsening grade.[2] The Gleason scoring system attempts to combine these diverse growth features into broad groups, primarily to promote

[*] Corresponding author at: 9000 Rockville Pike, Building 10, Room B3B69F, Artificial Intelligence Resource, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA.
E-mail address: stephanie.harmon@nih.gov (S. Harmon).

[1] Authors contributed equally.

http://dx.doi.org/10.1016/j.jpi.2024.100381
Received 6 November 2023; Received in revised form 27 March 2024; Accepted 29 April 2024
Available online 07 May 2024

reproducibility across pathologists but at the cost of reduced descriptive accuracy.[3] Pathologists assess prostate tumors by visually identifying the presence and prevalence of different architectural features, assigning a Gleason score and associated International Society of Urological Pathology system (ISUP) grade that aim to reflect the overall degree of tumor aggressiveness. However, due to the heterogeneity of morphologic features and the subjective nature of the process, there is poor inter- and intra-observer agreement in Gleason scoring,[4,5] which can easily over- or underestimate the actual tumor aggressiveness.

To address this challenge, several groups have developed AI-based approaches to build an automated and reproducible Gleason scoring algorithms.[6,7] However, most of these algorithms are designed to work with small-scale needle biopsies or tissue microarrays. Because each tumor contains multiple histological grades and each specimen can contain multiple areas of tumor across multiple slides, these algorithms may not reflect the tumor heterogeneity in whole prostate specimens.[8–11] Furthermore, very few approaches have been assessed for correlation with clinically relevant patient outcomes such as time to recurrence and development of metastasis, which requires detailed follow-up over a span of several decades.

A major limitation in using whole-mount prostate slides from surgical specimens in AI modeling is the presence of extensive inter- and intra-tumoral heterogeneity, which creates immense complexity for annotation[12] or, more often, limits these specimens to so-called “weakly labeled” annotations limited to a single index tumor or patient-level grades. Nagpal et al. overcame this issue by employing 29 different pathologists to obtain regional/pixel-level labels,[7] but such an effort is impractical over large multicenter cohorts. Other investigators have turned to data augmentation strategies to overcome labeling issues related to whole-image labels or tumor-based labels. For instance, CutMix is a popular data augmentation technique for improving classification performance and generalizability in tasks with localizable features.[13] It has specifically shown promise in digital pathology challenge sets where weakly labeled annotations produced from slide-level features do not localize exact features of disease.[14] Modifications of CutMix for histopathological classifiers in the setting of heterogeneous data labeling have also demonstrated that the generation of training sets from a combination of strong and weak labels, termed MixPatch, can improve performance and generalizability.[15] However, these applications have yet to be studied in prostate cancer histopathological grading.

In this study, we introduce a cascaded deep learning algorithm for cancer detection and Gleason pattern classification in digital pathology images of prostate specimens. Specifically, we aimed to investigate the effect of a selective CutMix training strategy involving real and artificially generated images from both biopsy and whole-mount specimens, on multi-class model generalizability despite heterogeneous pathologist annotations and annotation strategies. We additionally aimed to evaluate the prognostic value of AI-derived quantitative Gleason pattern estimates against pathologist-assigned Gleason scores when adjusting for known clinical risk factors.

Materials and methods

Patient populations and digital scanning

We used a multi-institutional collection of whole-slide images (WSIs) from biopsy and surgical specimens labeled by pathologists according to various annotation strategies. The datasets used in this study are summarized below for digital slide acquisition, digital annotation, and clinical data collection, including clinical follow-up when available. Inclusion criteria and annotation strategies are summarized in Fig. 1 and the total number of slides and Gleason score distribution utilized in this study are summarized in Table 1. Any data exclusions are additionally detailed in Fig. 1.

Dataset 1: The Prostate Cancer Grading Assessment (PANDA) public dataset consists of WSI from core needle biopsies obtained at the Radboud University Medical Center (RUMC) and the Karolinska Institute (KI).

- *Digital slide acquisition:* A total of $n = 10{,}616$ slides are available within these cohorts. WSI acquired at RUMC were scanned using a 3DHistech Pannoramic Flash II 250 scanner and shared publicly with a pixel resolution of 0.48 μm/pixel. WSI acquired at KI were scanned with either Hamamatsu C9600-12 or Aperio ScanScope AT2 scanners at 0.45202 μm/pixel and 0.5032 μm/pixel resolutions, respectively. Further information on how these annotations were acquired is available.[16]
- *Slide annotation:* Slides from Dataset 1 WSI were annotated separately at each center, where multiple pathologists provided either gland-level Gleason patterns or region-level Gleason scores. In the RUMC set, pixel-based annotations were provided at the gland level, with labels assigned according to background, stroma, benign epithelium, and cancerous epithelium labeled according to Gleason pattern (Gleason 3, 4, or 5). In the KI dataset, pixel-based regional annotations were provided for non-cancerous and cancerous regions. Representative images demonstrating the annotations are shown in Fig. 1B. Slide-level Gleason scores and ISUP grades were provided by both centers.
- *Clinical data and follow-up:* No additional clinical or pathological information was provided.

Dataset 2: Specimens from patients undergoing radical prostatectomy (RP) at the National Cancer Institute, enrolled in one or more of the clinical studies NCT03354416, NCT00102544, and NCT02594202 for clinical imaging and care of localized prostate cancer were retrospectively evaluated. Study inclusion was determined based on the availability of WSI, digital annotations by an expert pathologist, and clinical data.

- *Digital slide acquisition:* In total, 195 slides from 54 patients were available in digital format, acquired on one of two scanner types: Aperio (0.5404 μm/pixel) and Hamamatsu NanoZoomer (0.2212 μm/pixel).
- *Slide annotation:* Digital annotations were the result of re-review by an independent genitourinary (GU) pathologist to provide intratumoral region-level annotations corresponding to “pure” Gleason scores (3 + 3, 4 + 4, or 5 + 5) or regions of “true” mixed Gleason scores (3 + 4, 4 + 3, 4 + 5, or 5 + 4). Annotations were exported to JSON style format using QuPath software.17 A representative example of the annotations is provided in Fig. 1C.
- *Clinical data and follow-up:* Clinical data retrospectively collected included patient self-reported race, pre-RP prostate specific antigen level (PSA; ng/mL), age, and pathological findings collected from pathology reports including final patient Gleason grade, margin status, node status, stage, and various other histopathological features collected in accordance with clinical guidelines. Clinical follow-up from RP to the time of recurrence, defined either as biochemical recurrence based on serial PSA measurements or clinical evidence of recurrent and/or metastatic disease, was recorded.

Dataset 3: Archival tissues from RP specimens at the Joint Pathology Center were retrospectively retrieved for digital scanning and patients consented for research at the Center for Prostate Disease Research with follow-up data were selected for the study, under the protocol WRNMMC-EDO-2020-0657, 933668.

- *Digital slide acquisition:* At the time of study initiation, 398 slides from 138 patients' specimens were available for analysis. All slides were interpreted by a single expert pathologist. WSI were acquired on either Aperio (0.5016 μm/pixel) or 3D Histech (0.2425 μm/pixel) devices. All the provided WSI were unmarked, meaning only tissue was present on the slide.
- *Slide annotation:* In 138 patient specimen images acquired on the 3D Histech scanner, scout images of pathologist-inked markings from prospective evaluations were acquired prior to removal/cleaning of the ink and acquisition of full resolution WSI. These scout images were registered to WSI, and inked markings were digitally mapped back to WSI for downstream image processing (Fig. 1D). In these cases, the inked markings were used to provide cancer versus benign-level annotations and unique tumor lesion IDs. Pathologist-annotated tumor lesion IDs were used to link each tumor lesion with its assigned Gleason score from the clinical pathology reports. Tumor IDs are assigned by the pathologist at the

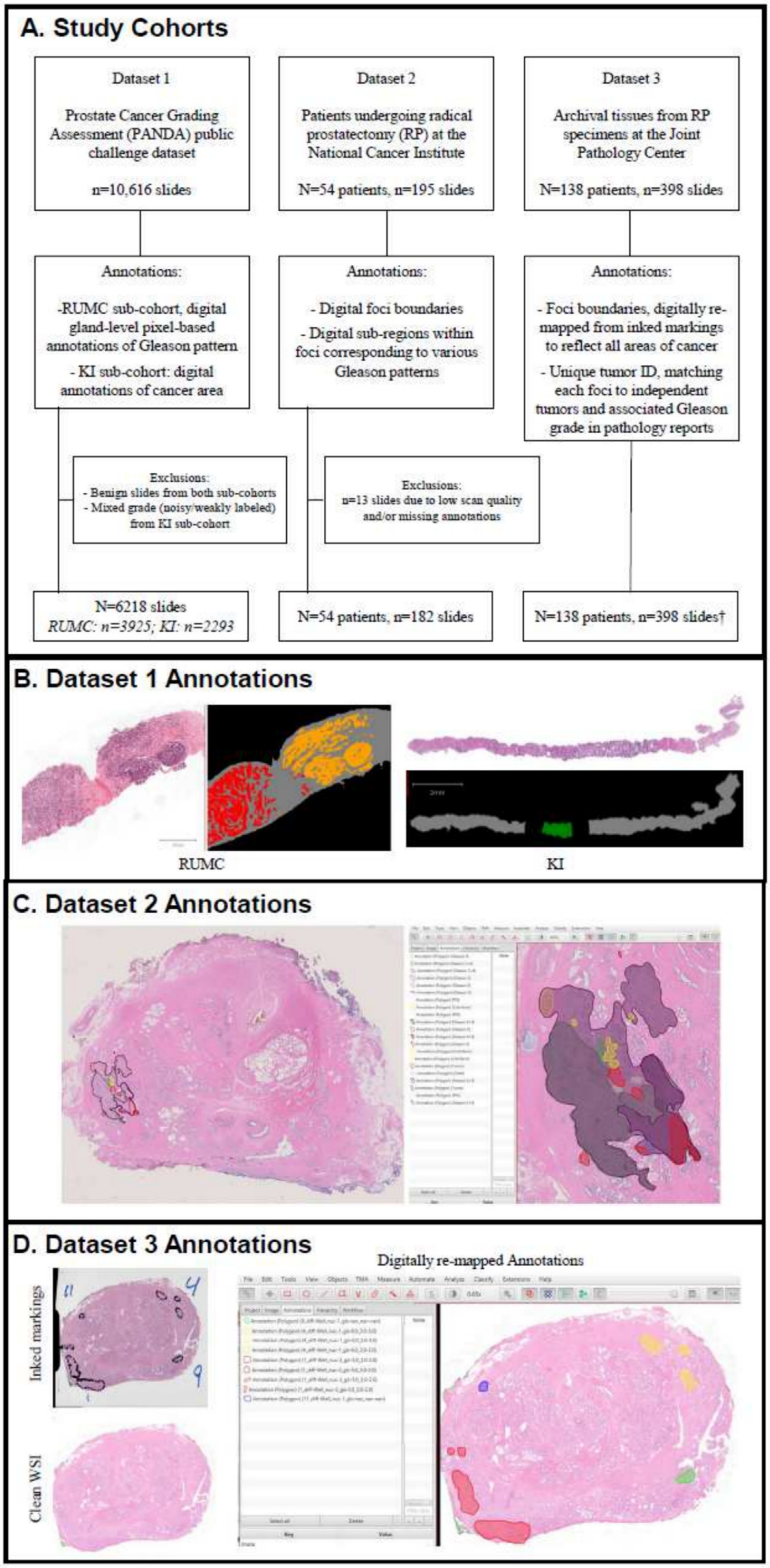
A. Study Cohorts
Dataset 1
Prostate Cancer Grading Assessment (PANDA) public challenge dataset
n=10,616 slides
Dataset 2
Patients undergoing radical prostatectomy (RP) at the National Cancer Institute
N=54 patients, n=195 slides
Dataset 3
Archival tissues from RP specimens at the Joint Pathology Center
N=138 patients, n=398 slides
Annotations:
-RUMC sub-cohort, digital gland-level pixel-based annotations of Gleason pattern
- KI sub-cohort: digital annotations of cancer area
Annotations:
- Digital foci boundaries
- Digital sub-regions within foci corresponding to various Gleason patterns
Annotations:
- Foci boundaries, digitally re-mapped from inked markings to reflect all areas of cancer
- Unique tumor ID, matching each foci to independent tumors and associated Gleason grade in pathology reports
Exclusions:
- Benign slides from both sub-cohorts
- Mixed grade (noisy/weakly labeled) from KI sub-cohort
Exclusions:
n=13 slides due to low scan quality and/or missing annotations
N=6218 slides
RUMC: n=3925; KI: n=2293
N=54 patients, n=182 slides
N=138 patients, n=398 slides†
B. Dataset 1 Annotations
RUMC
KI
C. Dataset 2 Annotations
D. Dataset 3 Annotations
Digitally re-mapped Annotations
Inked markings
11
4
9
1
Clean WSI

Fig. 1. Representative images of annotation strategy within each dataset. (A) Flow diagram outlining the number of patients/slides per dataset, annotation strategy, exclusions, and final number of patients/slides considered for analysis. (B) Dataset 1 *Left:* example from RUMC with detailed gland-level annotation from a slide containing Gleason patterns 4 (orange) + 5 (red). Background tissue is labeled as gray. *Right:* example from KI with regional annotations for background /benign in gray. (C) Dataset 2: example digital annotations completed within QuPath software, highlighting overall tumor region, and sub-regions enriched for specific patterns, including Gleason 5 (red), Gleason 4 (orange), Gleason 3 (purple), individual and mixed patterns (variable colors), and specific features such as cribriform (yellow) and PNI (blue). (D) Dataset 3 *Left-top*: example scout image reflecting foci ROIs drawn by a pathologist with inked markings and corresponding tumor identifier (tumor IDs, shown as numbers adjacent to foci markings) which are then mapped to tumor-specific Gleason grade from pathology report; *Left-bottom*: corresponding clean/unmarked whole-slide image; *Right*: digitally remapped foci segmentations and the corresponding tumor identifiers (labeled uniquely by color and corresponding grade from the pathology report). All individual foci corresponding to a tumor lesion are color-coded (4 = yellow, 9 = green, 1 = red, 11 = blue). The Gleason score assigned to a specific tumor lesion reflects the top two dominant Gleason growth patterns identified by the pathologist in that lesion. † $n = 30$ slides of benign tissue used for cancer detection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Total distribution of slides in each dataset according to patient-level Gleason score.

Gleason assignment		Dataset 1	Dataset 2	Dataset 3
Gleason score	Gleason sum			
3 + 3	6	2550	14	200
3 + 4	7	665	32	143
4 + 3		881	17	
3 + 5	8	62	0	12
4 + 4		1089	106	
5 + 3		39	0	
4 + 5	9	609	3	28
5 + 4		206	0	
5 + 5	10	117	10	9
Unknown				6
Total		6218	182	398

time of clinical interpretation (Fig. 1D) to identify regionally distinct tumors within all RP blocks. Each tumor is assigned an individual Gleason score within the pathology report, as well as an overall patient-level Gleason score. Of the 398 whole-mount prostate slides, 172 contained prospective annotations of the tumor lesions and the overall lesion-level Gleason score assigned to each tumor lesion.

- *Clinical data and follow-up:* Comprehensive pre- and post-surgical clinical and demographic data were collected as part of ongoing enrollment. Presurgical features included self-reported race, pre-RP PSA (ng/mL), and age. Surgical pathology reports were collected, including patient status (ISUP grade, overall Gleason grade [either Gleason score defined as primary + secondary patterns or Gleason sum],[6–10] margin status, node status, invasion, and pathological staging) and lesion status (tumor-specific Gleason score, nuclear grade, location, and organ confinement). Clinical follow-up from RP to the time of recurrence, defined either as biochemical recurrence based on serial PSA measurements or clinical evidence of recurrent and/or metastatic disease, was recorded. Clinical follow-up for the time from RP to detection and clinical confirmation of distant metastasis was also recorded. Demographics and clinical follow-up information for the cohort are presented in Table 2.

Table 2
Clinical data and demographics of Dataset 3 cohort.

Variable	Description	Summary
Gleason sum		
	6	54
	7	44
	8	8
	9	19
	10	4
	NA	3
Race		
	White	95
	African American	34
	Hispanic	3
	Asian	1
Age (years)	*Median (range)*	61.2 (40–75)
PSA (ng/mL)	*Median (range)*	6.04 (0.3–29.1)
Follow-up interval (years)	*Median (range)*	13 (<1–24)
Biochemical recurrence (years)		
	N events	45
	Time to event	1.45 (0.2–11.2)
Metastasis-free survival (years)		
	N events	23
	Time to event	6.58 (<1–17)
Overall survival (years)		
	N events	36
	Time to event	11.05 (2.5–21)

Image processing and data partitions

For all three datasets, tiles reflecting 500 × 500-pixel regions at an effective 10× magnification (1 μm/pixel) were extracted in a sliding window fashion from each WSI in the training and validation cohorts. Tiles were filtered to include a minimum of 5% tissue content. Tiles were labeled as benign or cancerous based on available pathologist labels (ground truths) per cohort for training the cancer detection algorithm, and cancer-containing tiles were labeled by Gleason score according to available ground truths, defined for each dataset.

Datasets were pseudo-randomly split into training, validation, and testing sets, with the aim of maintaining a similar distribution of patient-level Gleason scores across splits in each cohort.[18] Summary of cohort splits are provided in Supplemental Fig. 1. Dataset 1: As no clinical outcomes were available for this dataset, we primarily used it for training at a split at proportions 74/24/2 for train/validation/test. For the purposes of grading algorithm training, we included all cancer-containing cores from the RUMC set and selectively included “pure” grade (i.e., 3 + 3, 4 + 4, and 5 + 5) samples from KI to decrease label noise when training on tiles derived from each WSI. Dataset 2: Due to availability of highly detailed digital annotations and lack of long-term outcome data, this dataset was split in proportions 65/15/20 at the patient level. Dataset 3: Due to availability of treatment-related outcomes, this dataset was primarily used for testing and split at proportions of 18/7/75. In total, 318 slides from 104 patients were reserved for testing cascaded detection and grading algorithms. The distribution of Gleason scores assigned to the tissue specimens by pathologists at the time of clinical interpretation is shown in Table 1. Evaluation of algorithms in the test set was completed in sliding-window fashion for cascaded algorithms.

Cancer detection algorithm

To develop the algorithm for cancer detection, we used tiles to train a binary (benign versus cancer) classifier based on Mobilenet-V3 architecture with PyTorch. Tiles containing a minimum of 10% overlap with pathologist-defined cancer regions were labeled as “cancer,” with all others labeled as “benign.” An overlap threshold of 10% was chosen to ensure the cancer detection algorithm reliably detects all areas within the WSI containing tumor cells, including clinically relevant microscopic areas of the larger tumor delineation. Areas with <10% overlap to pathologist-defined tumor regions were unlikely to contain tumor cells and reflect majority benign tissue surrounding the foci. The model was implemented using Adam optimization, binary cross-entropy loss, and weights initialized from the ImageNet dataset. A cyclic learning rate starting at 1×10^{-5} was used, along with data augmentation strategies including MixUp,[19] contrast, and brightness variations to account for differences in hematoxylin and eosin (H&E) staining characteristics across various centers and scanners. After initial algorithm optimization, a random selection of cases from Dataset 3 training cohort were selected and WSI inference was completed. The binary outputs from cancer detection were converted to QuPath compatible geojson formats to enable the pathologist to load and interactively view the AI predictions (Supplemental Fig. 2). This process was used to determine the source of algorithm errors (false-positive- and false-negative predictions). An expert pathologist reviewed all AI predictions and interactively modified the ROIs using QuPath software. This process was completed in $n = 7$ cases, after which tiles were re-extracted from new ROIs and oversampled 5× in the training set. Models were re-trained and the final model was selected from the epoch with lowest validation loss.

Cancer grading algorithm

For training, each tile was associated with a one-hot encoded label vector y_i representing all possible Gleason patterns that may be present in the tile. For example, if a tile x_i overlaps with a 3 + 4 or 4 + 3 annotated tumor region, it was assigned a label $y_i = (1, 1, 0)$ representing the potential presence of Gleason patterns 3 and 4 within the tile. For each tile, the network outputs a vector of independent probabilities $p_i \in R^{3\times1}$, one per Gleason pattern, reflecting the probability of detecting each Gleason pattern. Due to regional heterogeneity, the true Gleason patterns present within a tile may not necessarily be reflected by the associated tile label, which captures the overall Gleason score of a tumor region. Therefore, tiles from regions broadly classified as having mixed grades, such as those from radical prostatectomy annotations or Karolinska subset of Dataset 1 (PANDA), were expected to have noisy labels. To mitigate overfitting to noisy labels, we used a hybrid training algorithm (Algorithm 1).

Algorithm 1. Selective Cut-Mix.

```
Input: $x_1, \ldots, x_n, y_1, \ldots, y_n, \alpha, \rho$.
Output: Multi-label Gleason pattern detector: $F(x_i, \theta) \rightarrow p_i \in R^{3\times1}$
#Initialize ResNet50 CNN using ImageNet pretrained weights;
$\theta \leftarrow \theta_0$;
for $epoch \leftarrow 1$ to $numepochs$ do
  for $batch \leftarrow 1$ to $numbatches$ do
    #pure grade examples;
    $X, Y \leftarrow (x_i, y_i)\ \forall i \in batch$ where $y_i \in \{(0,0,1), (0,1,0), (1,0,0), (0,0,0)\}$;
    #mixed grade examples;
    $X', Y' \leftarrow (x_i, y_i)\ \forall i \in batch$ where $y_i \in \{(1,0,1), (0,1,1), (1,1,0), (1)\}$;
    #randomly shuffle pure grade examples;
    $X'', Y'' \leftarrow$ RandomShuffle$(X, Y)$;
    #generate artificial examples with CutMix data augmentation of pure grade examples;
    $\tilde{X}, \lambda \leftarrow$ CutMix$((X, X''), \alpha)$;
    U(0, 1) ← random sample from uniform distribution
    if U(0, 1) > $\rho$ then
      #Compute cross entropy loss over real + artificial examples;
      $L_1 \leftarrow (1-\lambda)\,$BCELoss$(F(\tilde{X}, \theta), Y) + \lambda\,$BCELoss$(F(\tilde{X}, \theta), Y'')$;
      $L_2 \leftarrow$ BCELoss$(F(X', \theta), Y')$;
      $L \leftarrow L_1 + L_2$
    else
      #Compute cross entropy loss over real examples only;
      $L \leftarrow$ BCELoss$(F([X, X'], \theta), [Y, Y'])$;
    #error backpropagation;
    $\theta \leftarrow \theta - \eta \frac{\delta L}{\delta \theta}$
  end
end
```

The hybrid training algorithm uses the CutMix data augmentation technique[13] relies on artificial generation of "mixed" label tiles. In this application, consider an input pair of example images (x_i, x_j) from two separate pure Gleason labels (for example, 3 + 3 and 4 + 4). Our hybrid CutMix algorithm artificially derives a 3 + 4 tile by randomly cropping a region of x_i into x_j to generate a new example $\tilde{x}$ (Fig. 2). The size of the random crop is controlled by a size parameter $\lambda \sim Beta(\alpha, \alpha)$. Besides the learning rate η, the algorithm additionally takes as input a hyper-parameter ρ that controls, in a stochastic fashion, the extent to which we used artificially generated examples during training. For all purposes in this study, ρ was set to 0.5 to perform balanced training on real and artificial examples.

We used a ResNet50 convolutional neural network (CNN)[20] to perform multi-label Gleason pattern classification on stain-normalized WSI tiles $x_1, x_2, \ldots, x_i, \ldots, x_n$ from cancer regions. Stain normalization was done

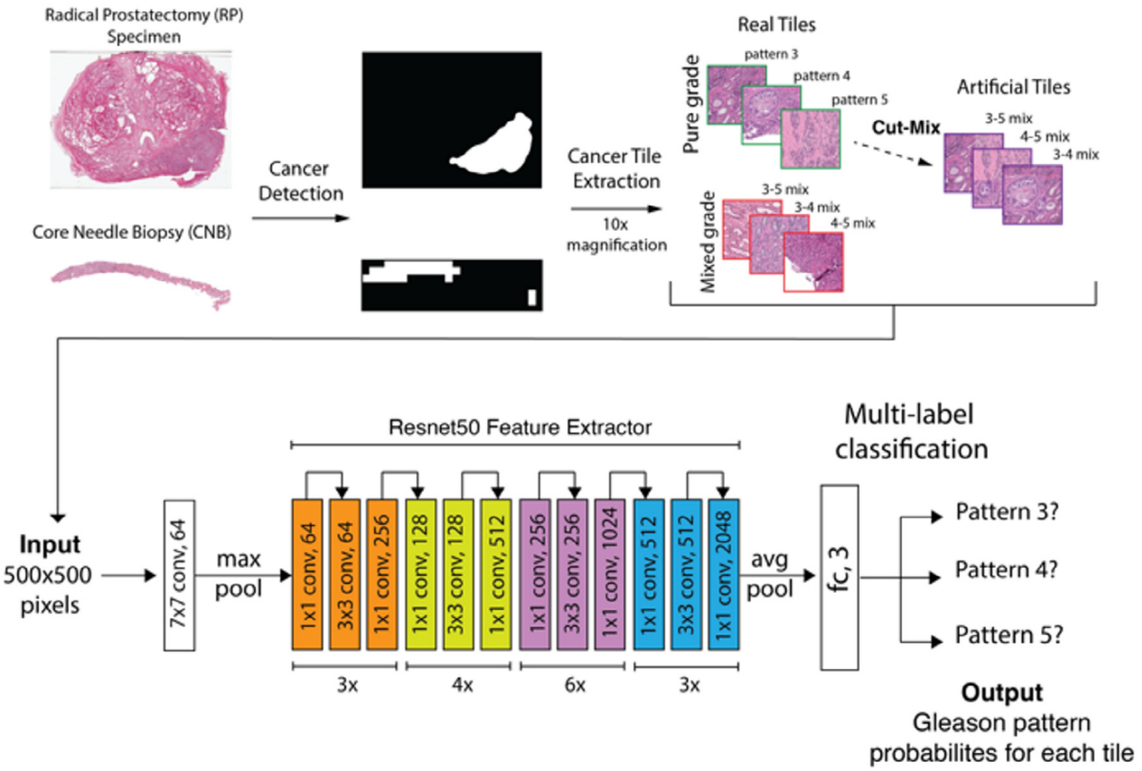

Fig. 2. Overview of the overall cascaded approach for Gleason grading. The first step of the algorithm involves cancer detection, which is performed using a trained MobileNetV3 classifier that learns to classify each whole-slide image tile as benign or containing cancer. In the second step, all cancer containing tiles are further analyzed using a multi-label classification algorithm to determine the probability of each Gleason growth pattern in each tile. The multi-label Gleason classifier consists of a Resnet50 feature extraction layer and a feed forward classification layer. The classification layer consists of three independent classification heads that predict the probability of observing each Gleason pattern. Training data consist of cancer tiles containing a mix of single and multiple Gleason growth patterns. To regularize model training, artificial tiles are generated using the CutMix data augmentation strategy,[13] which randomly cut-pastes regions of tiles from one class into tiles from another class to generate new artificial tiles containing a weighted mixture of image features from two classes.

using the Macenko method.[21] The cancer detection algorithm was used for pre-selection of patches for Gleason pattern classification, regardless of the presence of inked markings on the digital slides. In training and validation sets, tiles were re-extracted from false-positive regions within the cancer detection algorithm for training and validation and labeled as low risk, or Gleason 3, to avoid out-of-distribution errors in the grading algorithm. The learning rate η was initialized to 0.000001 and exponentially decayed over time using the exponential decay function, with the exponential decay parameter γ set to 0.9. The algorithm was implemented in Python using the PyTorch library. Back-propagation was performed using the Adam optimizer. The final model was selected from the epoch with the lowest validation loss and was frozen for inference on the test sets. A pictorial overview of the overall deep learning approach is depicted in Fig. 2.

To benchmark the performance of our proposed hybrid training strategy, we designed three negative control training experiments:

- Control 1 (Pure Grade + CutMix): Train a Gleason pattern classifier using tiles from pure-grade regions identified by pathologist annotation and artificially generated tiles derived using CutMix strategy as visually described in Fig. 2. Excluded tiles in this control include all mixed-grade regions by pathologist annotation.
- Control 2 (Pure Grade): Train a Gleason pattern classifier using tiles from pure-grade regions identified by pathologist annotation only. Excluded tiles in this control include all artificially generated tiles derived using CutMix strategy and all mixed-grade regions identified by pathologist annotation.
- Control 3 (Mix Grade): Train a Gleason pattern classifier using tiles from mixed-grade regions by pathologist annotation only. Excluded tiles in this control include all artificially generated tiles derived using CutMix strategy and all pure-grade regions identified by pathologist annotation.

Statistical analysis

Tumor detection performance was evaluated separately for each dataset due to variable reference annotations. In Datasets 2 and 3, detection sensitivity was reported per foci, where a tumor focus is defined as a unique spatial region of an individual slide. For Dataset 3, tumor-level sensitivity was additionally reported where multiple foci corresponded to a given tumor and a true-positive detection corresponded to positive detection of any foci corresponding to an individual tumor. Sorensen-dice similarity coefficient (DSC) was reported per slide to reflect the spatial correspondence of AI predictions and binary ground truth segmentations. DSC is defined as $2*(X \cap Y)/(|X|+|Y|)$, where X is the pathologist-defined ROI and Y is the AI-defined ROI. Detection performance by foci area (μm^2) was characterized in the testing set using the free-response operating characteristic (FROC) curve analysis to evaluate detection Sensitivity and number of FPs/image as a function of AI-predicted foci area.

To evaluate regional, tumor, and slide-level performance on the test set, spatial probability maps derived from a sliding-window inference were generated. Here, each pixel in the probability map reflects the AI-predicted likelihood for each Gleason pattern. Maps were derived for each training strategy (hybrid model and control experiments) within any area predicted as cancer-positive from the detection algorithm. Pixels containing >50% likelihood for any individual Gleason pattern were considered a positive prediction by AI. These maps were used to evaluate agreement with pathologist annotations, either using direct correlation to region-level annotations (Dataset 2) or by agreement with pathologist-reported slide-level (Dataset 1) or tumor-level (Dataset 3) grades. For agreement with region-level annotations (Fig. 1C), the proportion of AI-predicted regions containing each Gleason pattern within pathologist-defined ROIs were compared to the pathologist-assigned Gleason score using Kendall's Tau correlation accounting for clustered nature of data on the patient level was reported.[22] For agreement with tumor- and slide-level labels, we report the quadratic-weighted kappa metric for the exact Gleason score and ±1-score.[4] In Dataset 1, the total proportion of cancer-positive areas containing each Gleason score[3–5] predicted by AI were quantified for slide-level analysis. In Dataset 3, the total proportion of areas containing each Gleason score[3–5] predicted by AI were quantified within any ROI corresponding to each pathologist-identified tumor (Fig. 1D) for tumor-level analysis. In both settings, the AI-derived Gleason score sums are calculated by taking the sum over major and minor Gleason patterns present in the selected specimen regions and compared to the pathologist assignment.

To evaluate the correlation of AI-derived quantitative Gleason scores with patient outcomes, the weighted sum of the proportion of high-risk patterns 4 and 5 over all sections of the prostate were calculated within AI-derived tumor regions. We fit a stratified Cox proportional hazard model to the patients' clinical data to assess the prognostic power of quantitative Gleason scores while adjusting for positive or negative surgical margin assessment by the pathologist due to its strong prognostic correlation.[23] Outcomes analysis included RFS and MFS for Dataset 3. The Concordance Index (C-index) of these models is used to evaluate the discriminative ability of AI-derived quantitative Gleason scores across all experimental conditions (hybrid vs control experiments vs pathologist), it is analogous to standard AUC (while accounting for time-censored data) and a higher C-index is desirable. Partial likelihood ratio analysis for non-nested Cox-regression models was used to compare the C-index performance of AI-derived quantitative Gleason scores across all experimental conditions relative to ISUP. Finally, we performed unsupervised K-means clustering of tissue samples into low-, medium-, and high-risk groups based on AI-estimated features corresponding to the burden and proportions of each Gleason pattern within all slides samples for a given patient. The association of these groups with RFS and MFS was calculated and compared to pathologist-assigned Gleason Grade Groups (ISUP 1–2, 3, and 4–5).

Statistical analysis was completed in R (version 3.6.2.). Statistical significance was determined from the $p < 0.05$ level for all evaluations.

Results

Cancer detection performance on RP specimens

The initial algorithm trained with only Dataset 1 and used to evaluate generalization errors in RP specimens demonstrated 60.5% foci-level detection sensitivity and 86% tumor-level sensitivity, with average penalty of 2.98 $(0-31)$ false positives per slide. Training data from Datasets 2 and 3 were iteratively added, with interim results used for failure analysis from randomly selected training cases in Dataset 3, completed with pathologist review and assessment (Supplemental Fig. 2A). False positives were found to occur in atrophy, prostatic intraepithelial neoplasia (PIN), and periurethral tissue regions (Supplemental Fig. 2B). False negatives primarily occurred in small foci of low-grade cancer (Supplemental Fig. 2C). At the tile level, the final cancer detection algorithm achieved the best validation performance of 95.2% accuracy. The foci-level detection sensitivities were 78.6% and 92.7% in the Dataset 2 and Dataset 3 tests, respectively (Supplemental Table 1). Tumor-level detection sensitivity in Dataset 3 was 94.7%. DSC was higher in Dataset 2, with a median of 0.824 (range, 0.159–0.924) compared with Dataset 3, with median of 0.533 (range, 0–0.895). The mean number of false positives per slide was higher in Dataset 3 (13.3; range, 0–50) compared with Dataset 2 (2.48; range, 0–17). Foci-level FROC curve for the entire testing set is shown in Supplemental Fig. 3.

Gleason scoring performance

With respect to Gleason score performance, our hybrid learning algorithm achieved the lowest average validation cross-entropy loss compared with the three different controls when evaluated over 15 epochs (0.45 vs 0.50, 0.58, and 1.43 for controls 1–3, respectively). Due to the variation of data annotation within each dataset, statistical analysis was evaluated on region-, tumor-, and slide-level based on available data.

Region-level analysis was performed in Dataset 2. We demonstrated an improved ability to characterize regional intra-tumor heterogeneity of

Gleason scores within the sub-foci ROIs of distinct gleason patterns annotated by the pathologist in Dataset 2 (Fig. 3A–D). More specifically, based on Kendall's Tau correlation accounting for patient-level clustering, we observed an improvement in our correlation with proportion of pattern 5 across pathologist-assigned gleason patterns using the hybrid training approach compared with the 3 controls (Supplemental Table 2). Here, Gleason 3 is not expected to demonstrate a positive correlation as it is variably appearing in mixed grades, though overall decreases as expected in high-grade areas (Fig. 3). The extent of regional intra-tumor heterogeneity present within a single tumor is highlighted with a test case from Dataset 2 (Fig. 4), which had a tumor with an overall Gleason score of 4 + 4 along with tertiary Gleason 3 and Gleason 5 patterns. Fig. 4A depicts pathologist-marked region-level Gleason scores within one whole-mount slide, whereas Fig. 4B depicts pathologist-marked region-level Gleason scores for another whole-slide section of the same tumor.

Tumor-level agreement was performed in a subset of Dataset 3 for which tumor-level Gleason sum were available in the test set (n = 114 tumors from N = 28 patients). Table 3 reports the quadratic-weighted kappa metric based on AI-estimated proportions of major and minor Gleason patterns, demonstrating the hybrid training scheme resulted in a higher level of agreement with the pathologist. Contingency tables for per-class agreement are shown in Supplemental Fig. 4. For slide-level analysis, AI-estimated proportions within each biopsy slide were compared to ground-truth annotations from PANDA challenge (Dataset 1). Similar to tumor-level annotations, the hybrid training scheme resulted in a higher level of agreement with pathologist (Table 3), with majority of disagreement due to the AI-based prediction of pattern 4 presence (resulting in AI-predicted Gleason sum = 7) with pathologist-defined Gleason sum = 6 (Supplemental Fig. 5). Overall, the exact agreement with pathologists' assigned Gleason scores was higher for biopsy slides (κ = 0.63) compared with RP slides (κ = 0.44). This difference was not observed when accepting a Gleason ±1 score (Table 3).

Association of cascaded algorithms with patient outcomes

Finally, we assessed the prognostic value of our cascaded cancer detection and grading algorithms in test set specimens from Dataset 3 with matched RFS, MFS, and OS outcomes (N = 99 patients). Demographics and clinical follow-up information for the cohort are presented in Table 2. Overall, quantitative Gleason scores derived using the Hybrid training algorithm achieved a concordance index of 0.69 when modeling RFS, 0.72 when modeling MFS, and 0.64 when modeling OS times, outperforming the three different controls and traditional ISUP grading (Supplemental Table 3).

Unsupervised K-means clustering based on hybrid method AI-estimated proportions of each Gleason pattern demonstrated the optimal number of

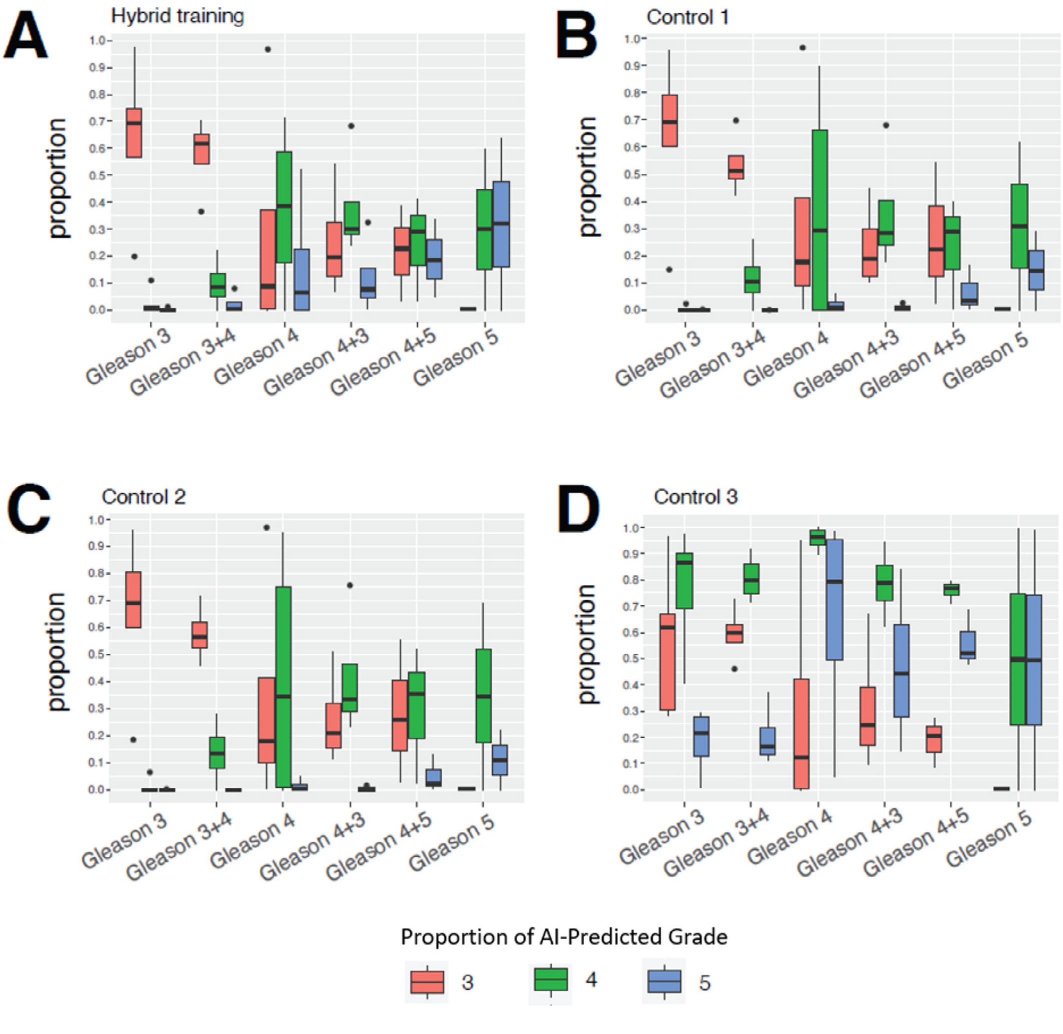

Fig. 3. (A–D) Boxplots depicting the distribution of the estimated proportion of each Gleason pattern in distinct tumor regions of all test cases from the NCI cohort (Dataset 2). X-axis: Pathologist-assigned Gleason score for any given tumor region. The pathologist-assigned Gleason score for a tumor region summarizes the top two dominant Gleason growth patterns detected by the pathologist in that region. Y axis: AI-estimated quantitative proportion of each Gleason growth pattern (depicted in red, green and blue). The proportion of a Gleason pattern in a tumor region is estimated as the fraction of tiles in that region that are predicted to have a >50% chance of containing that pattern. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

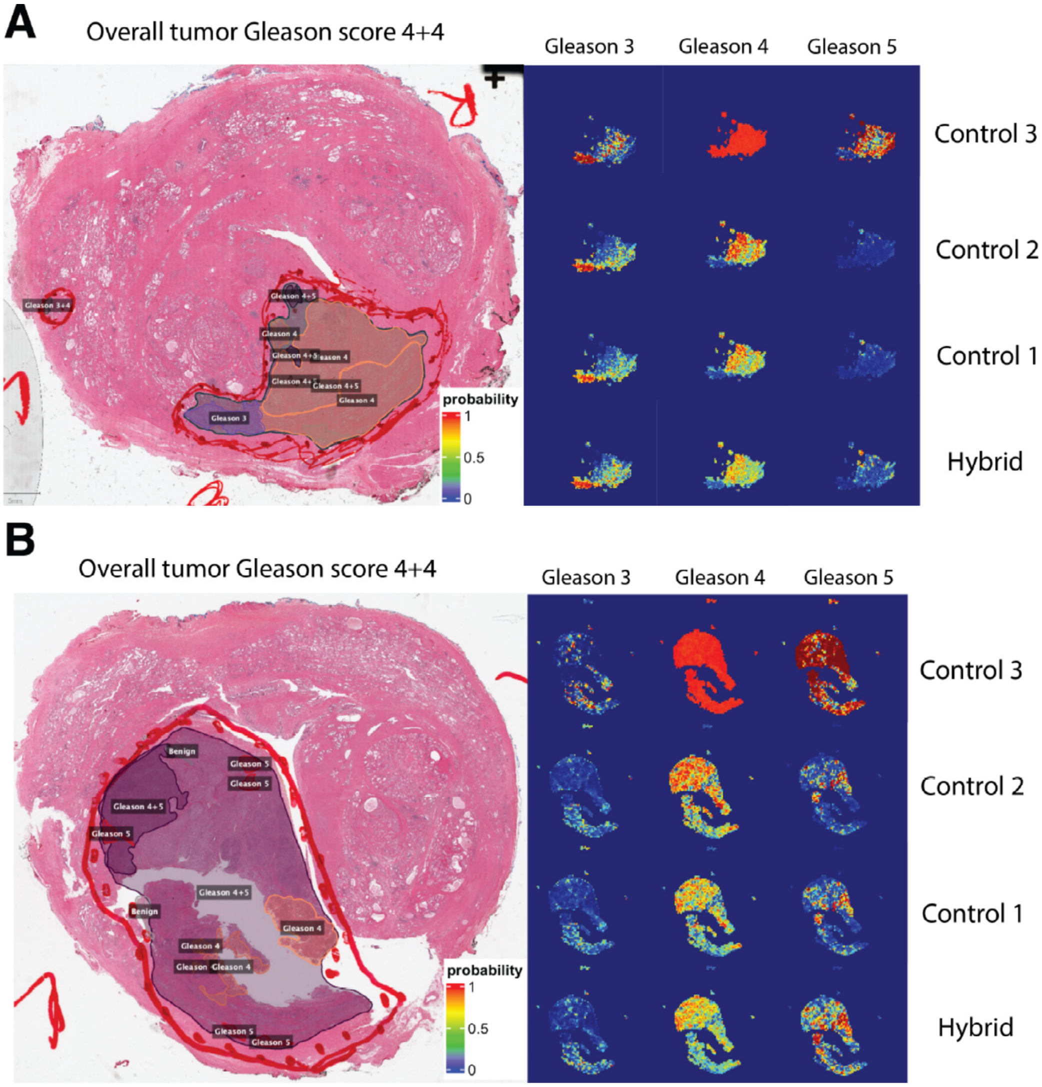

Fig. 4. (A–B) Depiction of region-level heterogeneity of Gleason growth patterns annotated by expert pathologist in two different sections of a test case from the NCI cohort (Dataset 2) with an overall case-level Gleason score of 4 + 4. Left: Pathologist-marked tumor regions with distinct Gleason scores. Right: AI-generated spatial probability maps depicting the spatial distribution of each Gleason pattern.

clusters to be based on silhouette scoring (Supplementary Fig. 6A–B). The resulting clusters demonstrate increasing proportions of Gleason 4 and 5 patterns based on AI predictions within whole-mount specimens (Supplementary Fig. 6C–D). This AI-based risk clustering revealed a significantly improved stratification of recurrence- and metastasis-free survival outcomes of patients in the test-set compared to ISUP grouping (Fig. 5). Cox proportion-hazard regression analysis demonstrated that medium- and high-risk samples from AI-based clustering indicated patients had significantly increasing risks of recurrence (medium-risk HR: 2.4 [1.2–5], p = 0.015; high-risk HR: 4.4 [1.4–14], p = 0.013) and metastasis (medium-risk HR: 3 [1.1–8.1], p = 0.03; high-risk HR: 20 [4.3–89.3], $p < 0.001$) compared with low risk group, even after accounting for known survival differences associated with surgical margins (Fig. 5). Correspondence of AI-based risk grouping and pathologist-based ISUP grouping is shown in Supplemental Fig. 7A. Further, Kaplan–Meier analysis reveals the likely source of improved outcome stratification is largely due to re-classification of ISUP 1–2 patients into medium- and high-risk AI groups for both MFS and RFS (Supplemental Fig. 7B–C). Within the limitations of the small cohort of ISUP 3–5 patients, improved stratification for ISUP 3 RFS and ISUP 4–5 MFS was also observed (Supplemental Fig. 7D–G).

Discussion

In this study, we introduce a cascaded deep learning algorithm to robustly detect different Gleason growth patterns within prostate tumor tissue acquired from needle biopsies or RP specimens. The main

Table 3
κ values for exact and ±1 unit agreement[4] between AI and pathologist-assigned Gleason score sum[6–10] for datasets with pathologist grading. Reference Gleason scores are defined as Dataset 1: pathologist grading of core needle biopsy (CNB) and Dataset 3: pathologist grading of distinct tumor lesions within RP specimens.

Dataset	Metric	Hybrid	Control 1	Control 2	Control 3
1 (CNB)	κ (exact)	0.63 [0.51–0.75]	0.42 [0.28–0.57]	0.5 [0.36–0.64]	0.3 [0.18–0.42]
	κ (±1 unit)	0.77 [0.65–0.89]	0.74 [0.60–0.87]	0.72 [0.59–0.86]	0.35 [0.22–0.48]
3 (RP)	κ (exact)	0.44 [0.15–0.73]	0.31 [0.11–0.73]	0.09 [0–0.26]	0.01 [0–0.03]
	κ (±1 unit)	0.79 [0.48–1]	0.56 [0.26–0.86]	0.29 [0–0.61]	0.01 [0–0.04]

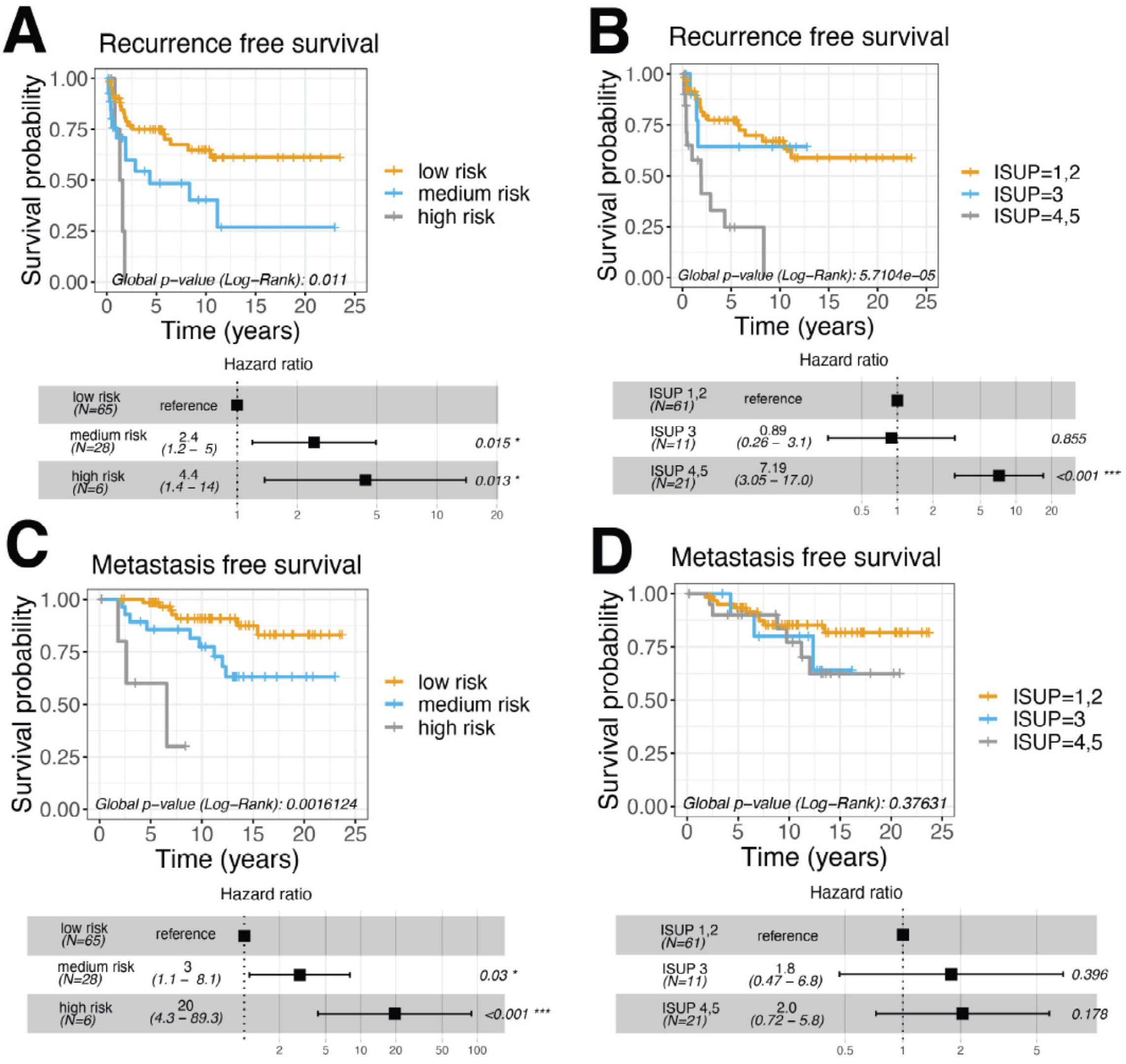

Fig. 5. (A–B) Kaplan–Meier plots and hazard ratios depicting stratification of recurrence free survival of test cases from the JPC cohort (Dataset 3). Patients were stratified into low, medium and high-risk groups based on AI-estimated quantitative proportions of each Gleason pattern (A) and pathologist assigned ISUP grades (B). (C–D) Kaplan–Meier plots and hazard ratios depicting stratification of metastasis free survival of test cases from the JPC cohort (Dataset 3). Patients were stratified into low-, intermediate-, and high-risk groups based on AI-estimated quantitative proportions of each Gleason pattern (C) and pathologist assigned ISUP grades (D). Hazard ratios and *p*-values were estimated by fitting a stratified Cox-proportional hazards model that accounts for survival differences associated with surgical margins. *: *p*-value <0.05, **: *p*-value <0.01, ***: *p*-value <0.001, ****: *p*-value <0.00001.

contributions of this work are twofold. First, we used a selective CutMix training strategy involving real and artificially generated images from both biopsy and whole-mount specimens, which improves multi-class model generalizability despite heterogeneous pathologist annotations and annotation strategies. Second, we demonstrated that AI-based quantitative grading improves the prediction of recurrence- and metastasis-free survival after RP. By harnessing detailed, long-term clinical follow-up data, we evaluate the prognostic value of AI-derived quantitative Gleason pattern estimations against pathologist-assigned ISUP grades, showing marked improvement in our ability to predict recurrence- and metastasis-free survival compared with traditional Gleason scoring, even after adjusting for survival differences associated with known clinical risk factors.

In contrast to previous approaches, ours uses a simple training strategy involving a combination of real and artificial training examples to impose heterogeneity on the training process. This not only improves the robustness and generalization of predictions in the face of interobserver variability but also allows us to characterize regional heterogeneity of Gleason patterns within individual tumors more precisely. Overall, the kappa values fell within the expected range of inter-rater agreement for needle biopsy and RP specimens.[5,24–26] Importantly, the agreement was the highest when using the hybrid training algorithm. Many of the disagreements between pathologist- and AI-derived Gleason scores arose from the higher resolution of AI predictions compared with pathologist annotations, especially in mixed-grade regions, where it is challenging for pathologists to define clear transitions between different Gleason patterns. A recent study demonstrates that pathologists' spatial annotations can vary up to 46% in size due to variations in annotation strategy and complexity.[27] These spatial variations, in addition to existing grading variations, can contribute to variable performance estimation in public-challenge datasets.[28] Taken together, our results suggest that a hybrid training strategy admixing heterogeneous labels improves the robustness and generalization of Gleason scoring despite errors associated with inter-observer variability and varied annotation styles.

We observed that the AI model's ability to capture subtle details in tumor regions resulted in improved stratification of RFS when using quantitative estimates compared with traditional scoring by pathologists. These results agree with a recent state-of-the-art study from Nagpal et al., who report a similar C-index, 0.65 (0.54–0.76), on whole-mount prostate specimens.[7] A follow-up study from the same group demonstrates that quantitative AI estimates of Gleason scores provide even better concordance with disease-specific survival (C-index 0.84), though their methodology, using leave-one-out cross-validation as opposed to an independent validation set, differed from that of previous studies and from our approach.[29] Other studies additionally support the notion that quantitative estimates of Gleason scoring and morphology characteristics may offer improved risk stratification for RFS.[30,31] Importantly, in contrast to recently reported studies, ours demonstrates a marked improvement in prediction of MFS over the long term using AI-derived Gleason scores, even after adjusting for known clinical confounding factors, including positive surgical margins.

Most prior models for assessing AI-based prostate cancer detection have been trained and validated in the biopsy setting.[9,11,16,32,33] We observed that when models are only trained using biopsy tissue, false positives occur in areas that are less commonly sampled during typical biopsy procedures, such as urethra, neural ganglion, or seminal vesicle tissue. Expert review of AI-predicted regions in a subset of the training population enabled faster identification and optimization of the final algorithm. Furthermore, we report variable region-level sensitivity depending on the level and quality of pathologist ground truth between datasets used in this study. In the case of digital annotations (Dataset 2), we report higher spatial correspondence, as measured by Dice coefficient, with the pathologist and lower sensitivity due to false negatives in small tumor foci regions. This is in comparison to performance against inked markings (Dataset 3), where we observed lower spatial correspondence in the Dice coefficient but higher sensitivity on both foci- and tumor-level annotations. This underscores the need for transparent reporting of annotation strategies, and further research on validation differences between methods is warranted.

Our study has several limitations. Intra-prostatic tumor regions, in resection or needle biopsy specimens, are often intermixed with stroma and normal prostate glands. Failure analysis of the cancer grading algorithm revealed that when the final cancer prediction area included benign areas adjacent to cancerous regions, the most likely grade assignment was Gleason pattern 3 with high probability by the AI grading algorithm. However, regions of stromal hyperplasia or immune cell infiltration were occasionally assigned as containing Gleason pattern 4 or Gleason pattern 5 by the AI grading algorithm with moderate probability (See Supplementary Fig. 8). All tissue areas were included in cancer detection evaluation, representing a strength of the sensitivity and low false-positive rate of the algorithm. However, cascading the detection and grading algorithms without user interaction could result in false-positive regions being assigned to high Gleason patterns. Further training and investigation of failure patterns is warranted. Similarly, while a diverse cohort of patient specimens was collected for algorithm training, it is still possible that histological variants such as neuroendocrine or squamous-like features are under-represented in the training set, and careful consideration of the performance of both detection and grading algorithms should be evaluated in future studies. Furthermore, there was a relatively limited representation of high risk disease within this cohort and further performance validation is warranted. Despite observed advantages of training from heterogeneous data samples, our study is limited by variable annotation strategies within each dataset which limit our ability to uniformly assess Gleason grading agreement with the pathologist across all datasets. A small number of cases were missing Gleason labels in Dataset 3, and therefore could not be statistically compared to pathologist grading. Each dataset was utilized differently depending on the nature of the provided annotations, resulting in a set of biased testing sets which could not be combined to one cohesive testing evaluation strategy. This warrants future validation in additional cohorts for evaluation on how our algorithm and training strategy may generalize to other clinical cohorts. Finally, prognostic validation of our quantitative cancer detection and grading algorithms was only performed in a relatively small cohort from a single institution, and therefore the generalizability of these findings to patient outcomes from other clinical or academic centers was not evaluated.

In summary, successful use of a hybrid CutMix approach that selectively combines data from strong and weak labels across multiple data sources results in generalizable performance across heterogeneous pathologist ground truths. Furthermore, our AI-based quantitative Gleason scoring approach has the potential to characterize tumor aggressiveness with greater precision compared to traditional Gleason scoring, and thus, improve prognostication after prostate surgery.

Disclaimers

The contents of this publication are the sole responsibility of the author (s) and do not necessarily reflect the views, opinions, or policies of Uniformed Services University of the Health Sciences (USUHS), The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., the Department of Defense (DoD), the Departments of the Army, Navy, or Air Force, or the U.S. Government. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government. The identification of specific products or scientific instrumentation is considered an integral part of the scientific endeavor and does not constitute endorsement or implied endorsement on the part of the author, DoD, or any component agency. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Ethics approval

Tissue specimens and corresponding clinical data were collected at each of the participating sites with IRB approval, which included sharing of de-identified data.

Author contributions

SAH, BT, PLC, JTM, IS, GC, GP, and AD were involved with study conceptualization. IS, RL, MM, GTB, DY, SE, JJ, KG, GC, SHT, FR, and JDM were involved with data curation, providing acquisition of clinical and imaging data. Methodology, software, data analysis and data visualization were completed by SAH and SP. SAH, SP, BT, PLC, AD, GP, IS, and JTM wrote and reviewed the manuscript. All authors read and approved the final submission.

Funding

This project is in part supported by the Intramural Research Program of the Center for Cancer Research, NCI. The project is additionally supported by the Center for Prostate Disease Research HU0001-20-2-0032 Co-Operative Agreement, between the Uniformed Services University and CPDR to G.T.C.

Data availability

Biopsy datasets from the PANDA challenge are publicly available (https://panda.grand-challenge.org/). A portion of Center 2 data is available in low resolution at the TCIA (https://wiki.cancerimagingarchive.net/display/Public/PROSTATE-MRI). At the time of submission, local IRB and ethics approvals were not obtained to allow public sharing of raw imaging data from all individual centers contributing to this paper. Readers are invited to contact the corresponding author for further information on data availability and data sharing policies. Code is available at: https://github.com/NIH-MIP/PCa-Detect-Grade

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The investigators used the computational resources of the NIH HPC Biowulf cluster. The authors thank Ellen Lazarus, MD, ELS for medical editing assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2024.100381.

References

1. Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep 1966;50(3): 125–128.
2. Aihara M, Wheeler TM, Ohori M, Scardino PT. Heterogeneity of prostate cancer in radical prostatectomy specimens. Urology 1994;43(1):60–66.discussion 66–7.
3. Epstein JI, Egevad L, Amin MB, et al. The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 2016;40(2):244–252.
4. Allsbrook Jr WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. Hum Pathol 2001;32(1):81–88.
5. Egevad L, Ahmad AS, Algaba F, et al. Standardization of Gleason grading among 337 European pathologists. Histopathology 2013;62(2):247–256.
6. Li WY, Li JY, Sarma KV, et al. Path R-CNN for prostate cancer diagnosis and Gleason grading of histological images. IEEE Trans Med Imaging 2019;38(4):945–954.
7. Nagpal K, Foote D, Liu Y, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. NPJ Digit Med 2019;2:48.
8. Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Sci Rep 2018;8(1), 12054.
9. 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;21(2): 233–241.
10. Singhal N, Soni S, Bonthu S, et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. Sci Rep 2022;12(1):3383.
11. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25(8): 1301–1309.
12. Duenweg SR, Brehler M, Bobholz SA, et al. Comparison ofmachine and deep learning models for automated tumor annotation on digitized whole slide prostate cancer histology. PLoS One 2023;18(3), e0278084.
13. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. CutMix: regularization strategy to train strong classifiers with localizable features. Proc IEEE Int Conf Comput Vis 2019:6022–6031.
14. Han C, Pan X, Yan L, et al. WSSS4LUAD: grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. arXiv preprint; 2022. arXiv:220406455.
15. Park Y, Kim M, Ashraf M, Ko YS, Yi MY. MixPatch: a new method for training histopathology image classifiers. Diagnostics 2022;12(6):1493.
16. Bulten W, Kartasalo K, Chen PC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med 2022;28(1):154–163.
17. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7(1), 16878.
18. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6(1).
19. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: beyond empirical risk minimization. arXiv preprint; 2017. arXiv:171009412.
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recognit 2016:770–778.
21. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: Proceedings of the Sixth IEEE International Conference on Symposium on Biomedical Imaging: From Nano to Macro; 2009; Boston, Massachusetts, USA. p. 1107–10.
22. Shih JH, Fay MP. Pearson's chi-square test and rank correlation inferences for clustered data. Biometrics 2017;73(3):822–834.
23. Karakiewicz PI, Eastham JA, Graefen M, et al. Prognostic impact of positive surgical margins in surgically treated prostate cancer: multi-institutional assessment of 5831 patients. Urology 2005;66(6):1245–1250.
24. Netto GJ, Eisenberger M, Epstein JI, Investigators TAXT. Interobserver variability in histologic evaluation of radical prostatectomy between central and local pathologists: findings of TAX 3501 multinational clinical trial. Urology 2011;77(5):1155–1160.
25. Persson J, Wilderäng U, Jiborn T, et al. Interobserver variability in the pathological assessment of radical prostatectomy specimens: findings of the laparoscopic prostatectomy robot open (LAPPRO) study. Scand J Urol 2014;48(2):160–167.
26. Veloso SG, Lima MF, Salles PG, Berenstein CK, Scalon JD, Bambirra EA. Interobserver agreement of Gleason score and modified Gleason score in needle biopsy and in surgical specimen of prostate cancer. Int Braz J Urol 2007;33(5):639–646.discussion 647–51.
27. Marrón-Esquivel JM, Duran-Lopez L, Linares-Barranco A, Dominguez-Morales JP. A comparative study of the inter-observer variability on Gleason grading against deep learning-based approaches for prostate cancer. Comput Biol Med 2023;159, 106856.
28. Nir G, Hor S, Karimi D, et al. Automatic grading of prostate cancer in digitized histopathology images: learning from multiple experts. Med Image Anal 2018;50:167–180.
29. Wulczyn E, Nagpal K, Symonds M, et al. Predicting prostate cancer specific-mortality with artificial intelligence-based Gleason grading. Commun Med 2021;1(1):10.
30. Serafin R, Koyuncu C, Xie W, et al. Nondestructive 3D pathology with analysis of nuclear features for prostate cancer risk assessment. J Pathol 2023;260(4):390–401.
31. Leo P, Chandramouli S, Farré X, et al. Computationally derived cribriform area index from prostate cancer hematoxylin and eosin images is associated with biochemical recurrence following radical prostatectomy and is Most prognostic in Gleason grade group 2. Eur Urol Focus 2021;7(4):722–732.
32. Li J, Li W, Sisk A, et al. A multi-resolution model for histopathology image classification and localization with multiple instance learning. Comput Biol Med 2021;131, 104253.
33. Pinckaers H, Bulten W, van der Laak J, Litjens G. Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels. IEEE Trans Med Imaging 2021;40(7):1817–1826.