


## Research Paper

# Radiologist-like artificial intelligence for grade group prediction of radical prostatectomy for reducing upgrading and downgrading from biopsy

Lizhi Shao<sup>1,2#</sup>, Ye Yan<sup>3#</sup>, Zhenyu Liu<sup>2,9,12#</sup>, Xiongjun Ye<sup>4#</sup>, Haizhui Xia<sup>3</sup>, Xuehua Zhu<sup>3</sup>, Yuting Zhang<sup>3</sup>, Zhiying Zhang<sup>3</sup>, Huiying Chen<sup>5</sup>, Wei He<sup>5</sup>, Cheng Liu<sup>3</sup>, Min Lu<sup>6</sup>, Yi Huang<sup>3</sup>, Lulin Ma<sup>3</sup>, Kai Sun<sup>2,8</sup>, Xuezhi Zhou<sup>2,8</sup>, Guanyu Yang<sup>1,7</sup>, Jian Lu<sup>3</sup> and Jie Tian<sup>2,8,10,11</sup>

1. School of Computer Science and Engineering, Southeast University, Nanjing, China.
2. CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging, the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
3. Department of Urology, Peking University Third Hospital, Beijing, China.
4. Urology and lithotripsy center, Peking University People's Hospital, Beijing, China.
5. Department of Radiology, Peking University Third Hospital, Beijing, China.
6. Department of Pathology, Peking University Third Hospital, Beijing, China.
7. LIST, Key Laboratory of Computer Network and Information Integration, Southeast University, Ministry of Education, Nanjing, China.
8. Engineering Research Center of Molecular and Neuro Imaging of Ministry of Education, School of Life Science and Technology, Xidian University, Xi'an, China.
9. CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
10. Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Medicine and Engineering, Beihang University, Beijing, China.
11. Key Laboratory of Big Data-Based Precision Medicine (Beihang University), Ministry of Industry and Information Technology, Beijing, China.
12. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100080, China.

#Co-first authors with equal contributions to this work.

 Corresponding authors: Prof. Guanyu Yang, School of Computer Science and Engineering, Southeast University, Nanjing, 211189, China. E-mail: yang.list@seu.edu.cn. Prof. Dr. Jian Lu, Department of Urology, Peking University Third Hospital, Beijing, 100191, China, E-mail: lujian@bjmu.edu.cn; Prof. Jie Tian, Fellow of AIMBE, IAMBE, ISMRM, IEEE, SPIE, OSA, IAPR, CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging, the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. (86)10-82618465, E-mail: jie.tian@ia.ac.cn.

© The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>). See <http://ivyspring.com/terms> for full terms and conditions.

Received: 2020.05.26; Accepted: 2020.08.21; Published: 2020.09.02

## Abstract

**Rationale:** To reduce upgrading and downgrading between needle biopsy (NB) and radical prostatectomy (RP) by predicting patient-level Gleason grade groups (GGs) of RP to avoid over- and under-treatment.

**Methods:** In this study, we retrospectively enrolled 575 patients from two medical institutions. All patients received prebiopsy magnetic resonance (MR) examinations, and pathological evaluations of NB and RP were available. A total of 12,708 slices of original male pelvic MR images (T2-weighted sequences with fat suppression, T2WI-FS) containing 5405 slices of prostate tissue, and 2,753 tumor annotations (only T2WI-FS were annotated using RP pathological sections as ground truth) were analyzed for the prediction of patient-level RP GGs. We present a prostate cancer (PCa) framework, PCa-GGNet, that mimics radiologist behavior based on deep reinforcement learning (DRL). We developed and validated it using a multi-center format.

**Results:** Accuracy (ACC) of our model outweighed NB results (0.815 [95% confidence interval (CI): 0.773-0.857] vs. 0.437 [95% CI: 0.335-0.539]). The PCa-GGNet scored higher (kappa value: 0.761) than NB (kappa value: 0.289). Our model significantly reduced the upgrading rate by 27.9% ( $P < 0.001$ ) and downgrading rate by 6.4% ( $P = 0.029$ ).

**Conclusions:** DRL using MRI can be applied to the prediction of patient-level RP GGs to reduce upgrading and downgrading from biopsy, potentially improving the clinical benefits of prostate cancer oncologic controls.

Key words: prostate cancer, Gleason grade groups, deep reinforcement learning, prostate cancer grading, magnetic resonance imaging



(TC2,  $N = 87$ , from PUPH) (Table S2). PC was used for model training. VC was used for internal verification. TC1 and TC2 were used for multi-center validation (Figure 1, Figure 2A).

**Table 1.** Patient characteristics

	PUTH	PUPH	<i>P</i> -value
Number of patients	488	87	
Age, median (IQR)	70 (65)	70 (64)	0.104
Total PSA, median (IQR)	11.5 (7.30)	14.7 (9.09)	0.596
Free PSA, median (IQR)	1.48 (0.87)	1.70 (0.89)	0.855
<b>Clinical T stage</b>			0.022
2a-2c	211 (43.2)	51 (58.6)	
3a	175 (35.9)	20 (23.0)	
3b	102 (20.9)	16 (18.4)	
Positive needle, median (IQR)	5 (2)	5 (3)	
<b>GG-NB, N (%)</b>			0.017
1	123 (25.2)	30 (34.5)	
2	98 (20.1)	23 (26.4)	
3	73 (15.0)	16 (18.4)	
4	99 (20.3)	11 (12.6)	
5	95 (19.5)	7 (8.0)	
<b>GG-RP, N (%)</b>			0.464
1	68 (13.9)	14 (16.1)	
2	126 (25.8)	26 (29.9)	
3	96 (19.7)	21 (24.1)	
4	75 (15.4)	10 (11.5)	
5	123 (25.2)	16 (18.4)	

Note: PUTH, Peking university third hospital; PUPH, Peking university people's hospital; GG-NB, grade group for pathological assessment of needle biopsy; GG-RP, grade group for pathological assessment of radical prostatectomy; *N*, the number of patients.

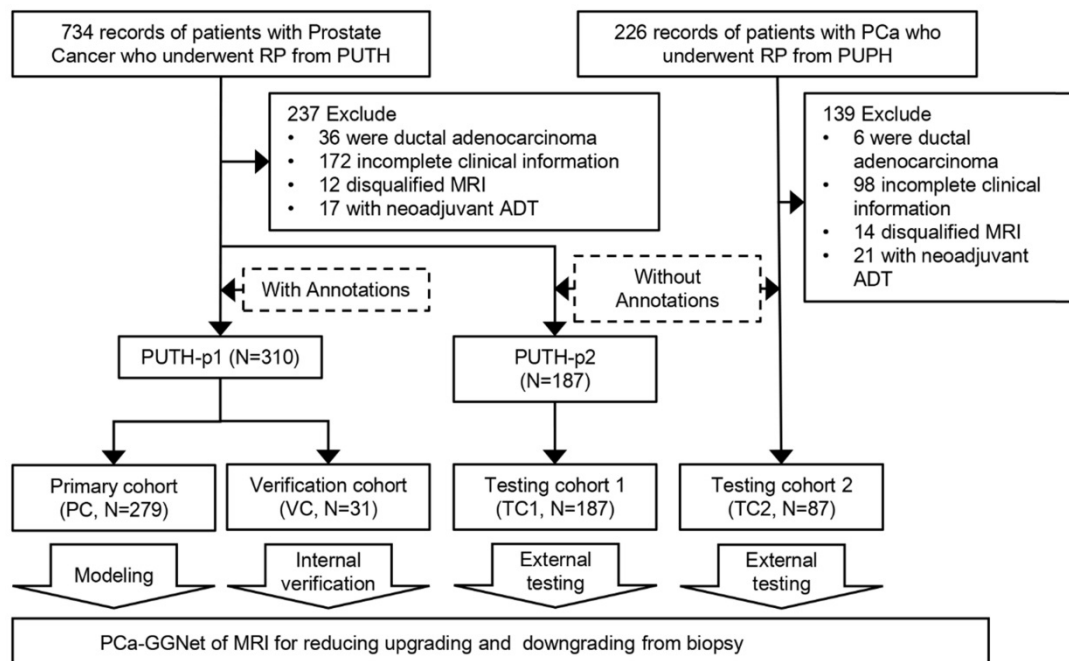
**Imaging data acquisition and annotations based on computational pathology registration**

All MRIs were performed before SBx using 3T MR scanners (Magnetom Trio, Siemens Healthcare,

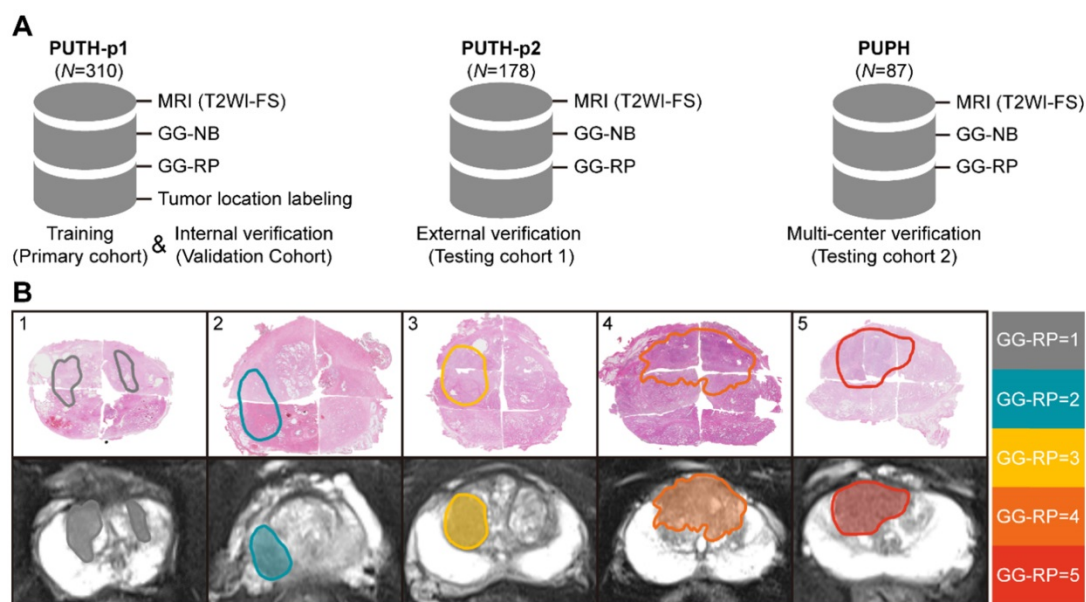
Erlangen, Germany/Discovery MR750, GE Healthcare, USA) without an endorectal coil. Only DICOM data of T2WI-FS (turbo-spin echo or fast-recovery fast-spin echo with fat suppression) were used for analysis in this study (Table S1).

Pathological hematoxylin-eosin sections of each patient from RP were scanned at 40× magnification to computational pathological sections (NanoZoomer S360, HAMAMATSU, Hamamatsu City, Japan). First, a pathologist having 22 years of urology expertise patched all the pieces into whole-mount sections and delineated the lesions that were responsible for diagnosis on each section. Second, our pathologist and one urological radiologist (12 years of experience) together recognized and delineated lesions on MRI correlated to whole-mount images, by using the knowledge of shape, texture, location of both the prostate and the tumors, which is knowing as cognitive registration. Of note, only lesions responsible for patient-level GG assessment were delineated. Very small satellite lesions or lesions contribute little to the final diagnosis were ignored. Five different examples are shown in Figure 2B (Supplementary Information I).

In our study, the T2WI-FS contained 24 [18-24] (Median [Min-Max]) slices per patient, in which 9 [8-12] (Median [Min-Max]) slices containing prostate gland were included. The number of annotations per case was 5 [4-10] (Median [Min-Max]). There was no significant difference between datasets in the distribution of the five-category GG-RP ( $P > 0.05$ ).



**Figure 1.** Patient recruitment and study design.



**Figure 2.** Dataset and annotations. **(A)** Multi-center datasets. **(B)** Tumor segmentation based on whole slide images of hematoxylin-eosin staining sections from different cases of RP.

### GSs/GGs of NB and RP

All patients received transrectal systematic biopsy at both centers. No targeted biopsy was performed. The number of cores ranged from 12 to 14. Subsequently, laparoscopic RP was performed one month after biopsy at either PUTH or PUPH. GS/GGs of NB and RP were reported at core, specimen, and patient levels according to the 2016 World Health Organization five-tier criteria [2]. Each pathology report was read and verified by two board-certified pathologists having PCa experiences of 6 and 22 years, independently. For cases having different assessments, a thorough discussion was conducted to reach a final agreement (**Figure S1**).

### Deep CNN for slice-level analysis using identified tumor slices

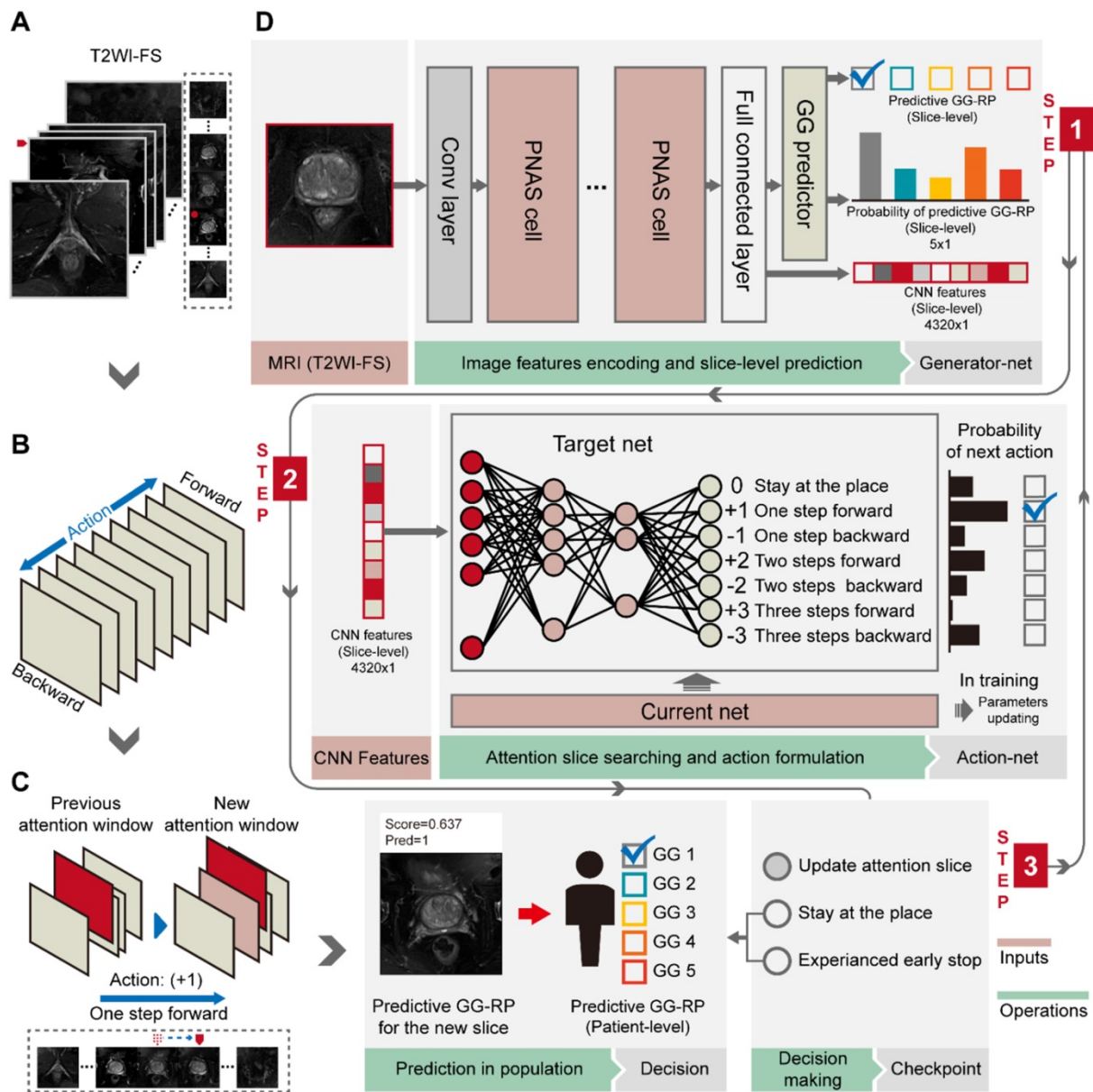
First, we performed a pixel-wise analysis to obtain slice-level prediction and CNN features using tumor slices of T2WI-FS (**Figure 3A, 3D**). A PNASNet-5-large [24]-based progressive search strategy was adopted as the structure for constructing a classification model (generator-net), which earned state-of-art performance for image classification with an accuracy of 1,000-category on the ImageNet [25] test set: 82.9% (top-1) and 96.2% (top-5). Model parameters of the model trained by ImageNet were used for the pre-training network and for transfer learning [26], in which the filter parameters of the network were frozen, except for the last five layers. Next, the model was trained with semi-supervised learning, and the label of each slice was consistent with patient-level GGs. During model training, data

augmentation was used to restrict overfitting, including random rotation, mirror transformation, and affine transformation. The central point of the original image window was the anchor point and the area with a window size of  $200 \times 200$  (pixels  $\times$  pixels) near the anchor point was selected as the region of interest (ROI) to focus the network's attention on the prostate area. The ROI was then scaled to  $331 \times 331$  (pixels  $\times$  pixels) via bilinear interpolation as input.

Additionally, inputs were converted into a three-channel image, and each channel was standardized with a mean of 0.5 and a variance of 0.5. CNN features related to PCa GGs were then extracted from the last fully connected layers, providing a vector ( $f$ ) of  $4,320 \times 1$ . The output of the classifier was generated by softmax, having a vector ( $k$ ) of  $5 \times 1$ , including the corresponding prediction probability of the five-grade GG,  $p_{i,k}$ . The category having maximum prediction probability contained the predicted GGs. During training, the batch size of every interaction was 64, and the loss function was defined by the cross-entropy of multiple classifications to update filters via backpropagation (**F-1**). The consistencies between labels and predicted results were binary ( $y_{i,k}$ ). The learning rate was set to 0.01 with an exponential reduction of 0.97, and the momentum was adjusted at 0.9. When the epoch training finished, the VC was employed for internal validation and early stopping to prevent overfitting. The model training was terminated and saved until the overall ACC of five consecutive epochs was stagnant, giving us the generator net.

$$L_{\log}(Y, P) = -\log \Pr(Y|P) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} \log p_{i,k} \quad (\mathbf{F-1})$$





**Figure 3.** Workflow of PCa-GGNet. **(A)** The input of PCa-GGNet, for which only one slice was input per operation. The initial input was the median slice of the whole T2WI-FS sequence. **(B)** Action rules for attentional slice searching, which included direction and step length of actions. **(C)** The illustration of attentional slice searching and updating. **(D)** The workflow and architecture of PCA-GGNet. PNAS refers to a progressive neural architecture search. In the first step, we selected the median slice of T2WI-FS as input for the convolutional neural network (CNN)-based model to predict GG-RP on each slice. For the second step, we used features from the first step of the DRL-based model to generate an action for updating input. For the third step, a checkpoint was used to determine whether the results of the current input could be used as patient-level predictions. If so, the input was a decision slice. If not, our attention was changed through actions to find a new slice as an input. The framework was a sequential method to predict patient-level GG-RP. The gray arrow shows a forecasting process. See Methods for complete details.

### DRL for simulating radiologist reading behavior to search for attentional slices

We used a DRL strategy to mimic the diagnostic behavior of radiologists and to improve slice-level-to-patient-level prediction. The model slid each slice of the 3D T2WI-FS forward or backward using an attention mechanism. It then associated the memory of the browsing path to empower the attentional slice searching strategy. A slice was finally obtained as the decision slice, identifying the most critical slice for the patient-level GG-RP. For the experiments, we adopted

the deep-Q network (DQN) [22], which has been used to replicate the human-level player performance in sports video games, as the basic structure. We redesigned the game mechanism and reward function for the GG prediction problem. The DQN consists of current and target nets having the same configuration as an artificial neural network with two hidden layers. The input of net ( $s$ ) is a  $4,320 \times 1$  vector (denoted as the status), and the output ( $Q$ ) is a  $7 \times 1$  vector that indicates different orders of action ( $a$ ) (Figure 3B, 3D). The two hidden layers were constructed with 50 and 30 neurons, respectively. The current net was used to

collect experiences into a pool during training and to update their parameters using  $Q\_loss$  (F-2). The collected experience included the rewards,  $r$ , underlying the current status, and actions. Rewards are defined by their predicted probability,  $P_{s,a}$ , and the consistency between the predicted and true labels (F-3). The basic reward ( $y_{s,a}$ ) and the reward rate ( $\alpha$ ) of predicted probability ( $P_{s,a}$ ) were set as 1 and 0.5. When the experience pool overflowed, the benefits of a single action in the experience pool were randomly recorded, and the neuron parameters of the current net were assigned to the target net when the number of accumulations reached 100. The training environment included patients who provided 3D T2WI-FS slice imagery with identified CNN features. We changed the training environment by randomly selecting the starting slice to achieve data augmentation, which increased the robustness of the model. During the attentional slice searching phase, only the target net was used as a decision agent to determine the probability of actions and to select the action having the highest probability.

$$Q\_loss = E[(Target_Q - Current_Q)^2] = E[(r + \gamma \max_{a'} Q(s', a'; \theta) - Q(s, a; \theta))^2] \quad (\text{F-2})$$

$$r = y_{s,a}(b + \alpha y_{s,a} P_{s,a}) \quad (\text{F-3})$$

### Patient-level prediction of PCa-GGNet framework

To construct a GG prediction indicator at the patient-level, a three-stage PCa-GGNet framework was developed. Two basic units (i.e., generator and action net) were prepared during the training phase. A slice was the framework's input. A classifier based on the tumor slice was established for five-category prediction at the slice-level (i.e., generator net) (Figure 3D). Inputs for the training generator net included tumor slices, and their labels reflected the patient-level GG-RP. Next, the action net was trained for attentional slice searching using features and classification results, as generated by the generator net. To train the action net, we defined the T2WI-FS slices as the environment, for which labels included patient-level GG-RP and flags of tumor slices. The generator net and action net were built step-by-step. During the prediction phase, three steps were required for the PCa-GGNet framework to predict patient-level GG. In this first step, the middle slice was selected as the input to the generator net (Figure 3A), and CNN features and predictions based on slices were generated from the generator net. During the second step, CNN features were employed for the action net to produce an action order based on rules (Figure 3B). Lastly, a checkpoint was set to draw a conclusion based on the action order from the action net. If the framework-running circle was not satisfied

with the condition of the checkpoint, which would experience an early stop or stay-in-place action, the framework would update the current attentional slice and repeat steps 1 and 2 (Figure 3C). Otherwise, the patient-level prediction would adopt the result of the attentional slice generated from the last circle. The initial input was the median slices, which recorded the radiological information of the prostate area.

### Evaluation

Quantitative statistics were summarized as mean  $\pm$  standard deviation. Categorical variables were achieved via the  $\chi^2$  test or Fisher's test. The reported statistical significance levels were all two-sided, with the statistical significance level set to 0.05. ACC and quadratic Cohen's kappa coefficient were used to evaluate the overall performance of the multi-category classification. Precision (F-5), recall (F-6), and F1-score (F-7) were used for evaluation within the category. The 95% confidence interval (CI) values were calculated using a bootstrap strategy ( $N = 1,000$ ). Statistical analyses were performed using Python's (v.3.6.5) scikit-learn package (v.0.21.3) and R (v.3.1.0).

$$\text{Precision} = \frac{\text{True positive}}{(\text{True positive} + \text{False positive})} \quad (\text{F-5})$$

$$\text{Recall} = \frac{\text{True positive}}{(\text{True positive} + \text{False negative})} \quad (\text{F-6})$$

$$\text{F1 - score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})} \quad (\text{F-7})$$

## Results

### Inconsistency between biopsy assessment and RP pathology

The consistencies between the Cohen's kappa values of GG-NB and GG-RP were 0.364 and 0.289 at PUTH and PUPH, respectively. The mean accuracy (ACC) between GG-NB and GG-RP on the total patients was 0.484 (95% CI: 0.379-0.588). The overall upgrading rate reached 40.4%, which was significantly higher than the downgrading rate of 14.7% ( $P < 0.001$ ) (Figure 4). The upgrading and downgrading of each GG-NB are shown on the left side of Figure 5. Apart from GG 1, the second-largest proportion of upgrading was in GG 3, and the largest downgrading cases were in GG 4. Importantly, more than 50% of GG-NB 3 upgrades shifted to GG-RP 5, and some patients in GG-NB 4 or 5 downgraded to 3 or lower. Consistency analysis of GGs between GG-NB and GG-RP is shown in Table S3 for different cohorts.

### Assessment of generator net for slice-level GG-RP prediction based on lesion slice analysis

To construct a multi-classification model for predicting patient-level GG-RP, we first built a five-

category prediction model (i.e., generator net) based on lesion-level analysis to distinguish different GG-RPs as accurately as possible. During the training phase, a total of 1,484 T2WI-FS tumor slices in PC were used for the model's parametric learning. A total of 160 tumor slices from VC were used as internal verification and were regarded as the model's early terminating conditions to prevent overfitting.

The ACC of generator net for the five-category (GG-RP 1-5) classification in PC and VC were 0.73 (95% CI: 0.711-0.749) and 0.615 (95% CI: 0.545-0.686), respectively (Table S4). More details (i.e., precision, recall, and F1-score) at each grade are listed in Table S5. We also tried to merge slices without tumor annotations into original samples as a separate category to construct a six-category model to not only distinguish five levels of GG-RP, but also to filter-out slices without tumors. Although the overall ACC of the classifier was improved to 0.838 (95% CI: 0.83-0.847) in PC and 0.803 (95% CI: 0.777-0.829) in VC, the ACC for GG-RP prediction (slice-level) significantly dropped to 0.54 in PC (95% CI: 0.517-0.562) and 0.523 (95% CI: 0.451-0.594) in VC (Table S4), respectively. Therefore, the five-category model containing tumor annotations of training samples was adopted as the generator net for predicting GG-RP at the slice-level. We also compared the classification performance of different network structures in Table S4, and the most optimal basic network structure was the PNASNet-5-large net.

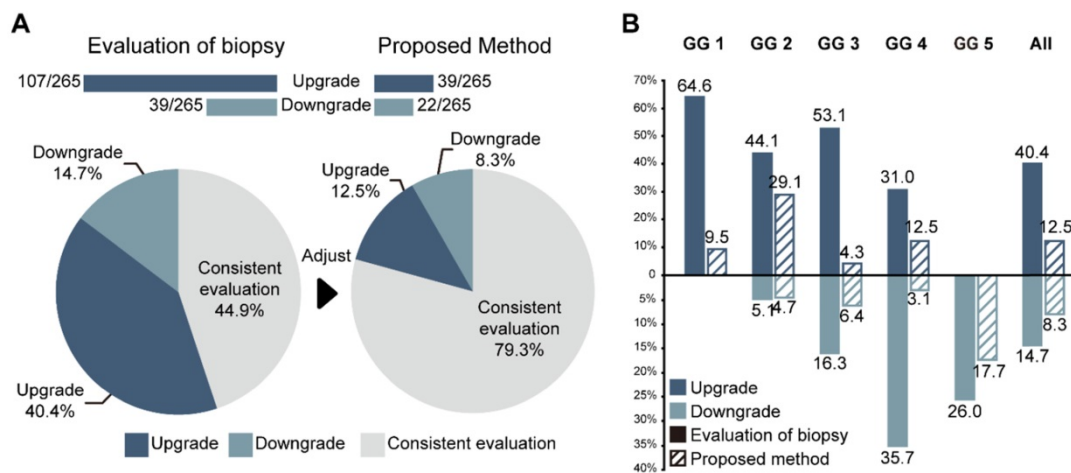
### Performance of discriminator net for attentional slice searching

Based on features from the generator net output, the action net was modeled for attentional slice searching to update the input of the generator net in a new prediction period and to draw the final decision. The ACC of the action net (designed to identify slices

containing tumors in the 3D T2WI-FS) was 0.862 (95% CI: 0.85-0.874) in PC by five-fold cross-validation (Table S6). According to the rules, no matter whether if we received a "stay at the place" status at the last step, we adopted the last-searched slice as the attentional slice, so that the model would keep the sensitivity of 100%. The specificity of the model was 0.86 (95% CI: 0.848-0.872) on PC. In the experimental attempts, the four-circle was set as the terminating condition of the action net, in which the slice at the fourth act was used as the basis for the final decision. Next, we verified the action net in VC with an ACC of 0.797 (95% CI: 0.754-0.841) and a specificity of 0.797 (95% CI: 0.754-0.841) (Table S6). The ACC of the GG-slice to the finally selected slices was 0.86 (95% CI: 0.846-0.874) in PC and 0.832 (95% CI: 0.784-0.88) in VC (Table S6).

### Assessment of PCa-GGNet for predicting GG-RP at patient-level and restriction of upgrading and downgrading risks

To explore whether the PCa-GGNet using T2WI could construct a prediction index highly related to the GG-RP at the patient-level, we first constructed a computing framework and trained it in PC. The prediction process is visualized in Figure S2. The predicted GG from PCa-GGNet (GG-Pre) obtained a five-category ACC of 0.847 (95% CI: 0.826-0.867) in PC and 0.83 (95% CI: 0.762-0.898) in VC (Table S3). The kappa consistency between the GG-Pre and GG-RP was 0.804 (95% CI: 0.752-0.857) and 0.777 (95% CI: 0.599-0.954) in PC and VC, respectively. The F1-score of each GG (1-5) was 0.893 (95% CI: 0.767-1.018), 0.79 (95% CI: 0.697-0.883), 0.62 (95% CI: 0.341-0.899), 0.868 (95% CI: 0.725-1.012), and 0.877 (95% CI: 0.732-1.022), respectively (Table S7). Furthermore, to validate the reliability of the PCa-GGNet, the model was tested on multi-center datasets obtaining ACCs of 0.781 (95%



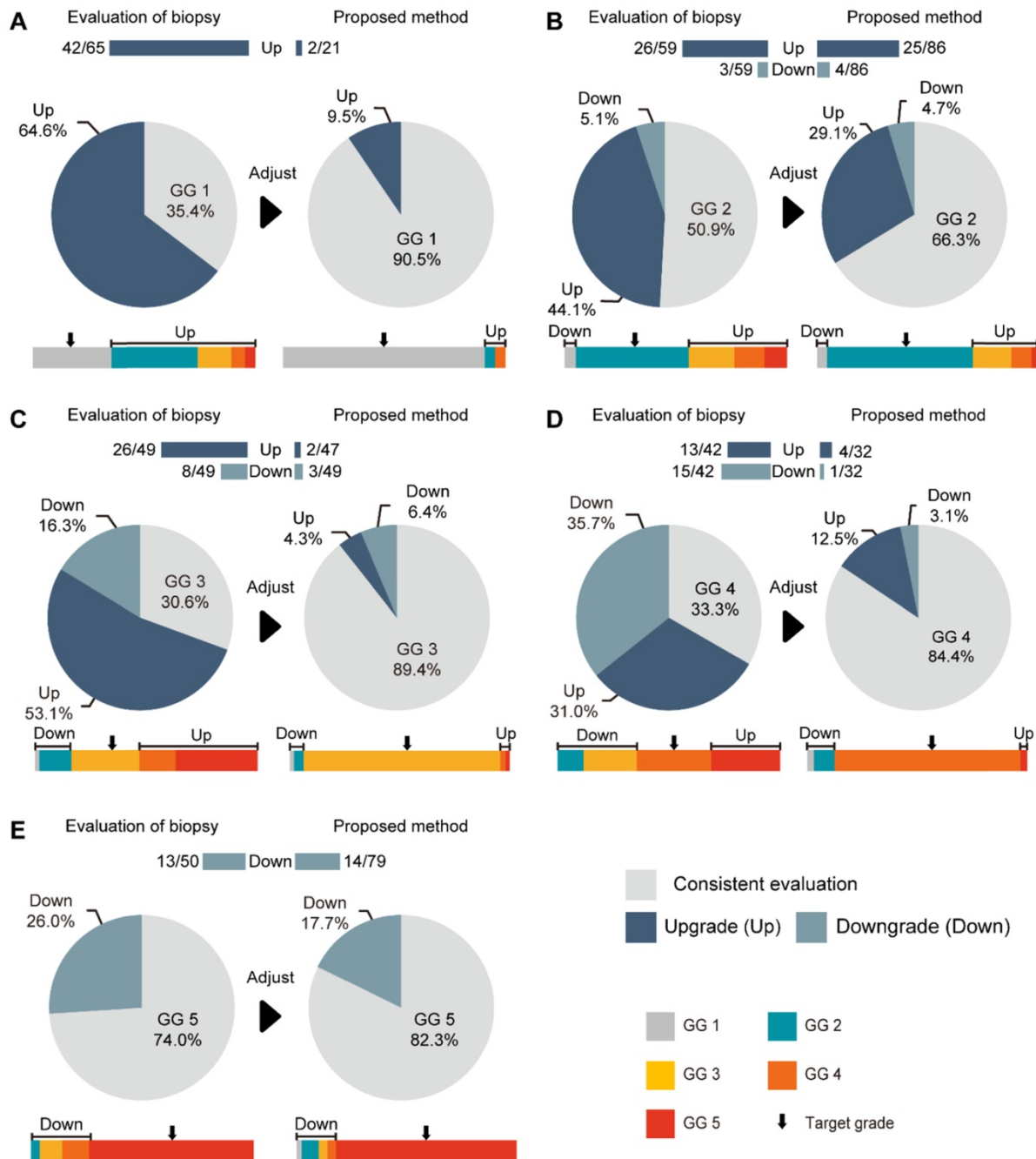
**Figure 4.** Performance of PCa-GGNet against GG-NB for upgrading or downgrading evaluation. (A) Overall performance. The bar chart and ratio at the top indicate the number of people upgraded or downgraded in the dataset. (B) Comparison of the rates of upgrading or downgrading between biopsy and our method.



CI: 0.751-0.811) in TC1, and 0.815 (95% CI: 0.773-0.857) in TC2 (Table S3). The kappa consistency in TC1 was 0.713 (95% CI: 0.632-0.794) and 0.761 (95% CI: 0.656-0.865) and TC2, respectively. The confusion matrix between GG-Pre and GG-RP under multi-center settings are shown in Figure 6E-H. ROC analysis was used for evaluating the performance of GG-Pre according to different subgroups, AUCs of low-grade (grade 1 vs. 2,3,4,5), medium-grade (grade 1,2 vs. 3,4,5) and high-grade (grade 1,2,3 vs. 4,5) groups were all greater than 0.8 in PC, TC1, and TC2

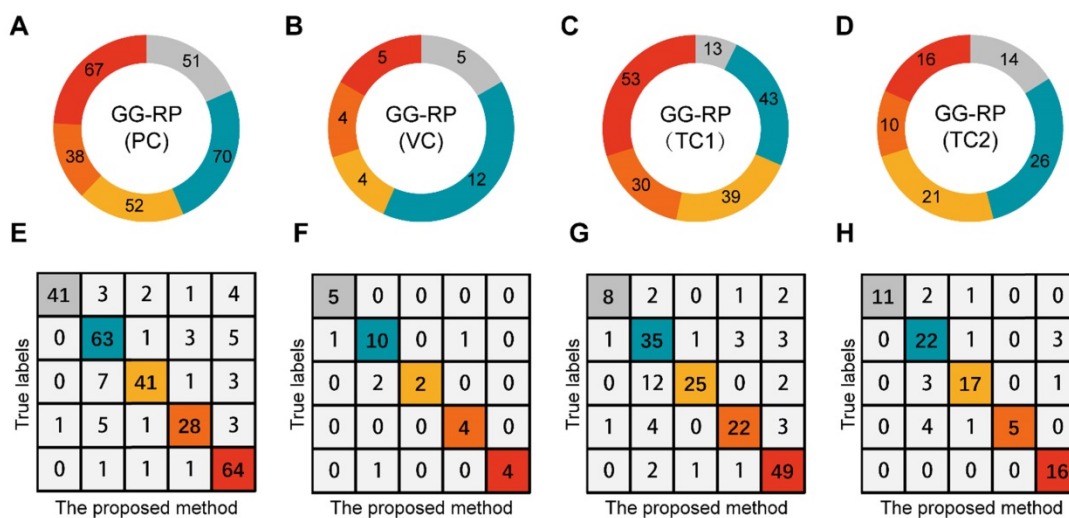
(Figure S3).

For assessing the restriction of upgrading and downgrading risks, the inconsistency of PCa-GGNet in all testing samples decreased to 12.5% (upgrading) and 6.3% (downgrading) (Figure 4A). Consistency ratios of PCa-GGNet at each GG (1-5) were 90.5%, 66.3%, 89.4%, 84.4%, and 82.3%, respectively (Figure 5). Top-2 predictions of PCa-GGNet were in grade 1 with an F1-score of 0.876 (95% CI: 0.805-0.947) and in grade 5 with an F1-score of 0.884 (95% CI: 0.826-0.942) (Table S7).



**Figure 5.** Comparison between our method and GG-NB at each grade. (A) Grade 1 of GG-RP. (B) Grade 2 of GG-RP. (C) Grade 3 of GG-RP. (D) Grade 4 of GG-RP. (E) Grade 5 of GG-RP. The upper bar chart and ratio at the top indicate the number of people who were upgraded or downgraded in the dataset. The color bar represents the detailed type of upgrading and downgrading group. The black arrow indicates the target of the GG-RP corresponding to the pie chart.





**Figure 6.** Assessment of PCa-GGNet in multi-center for multi-category prediction. **(A-D)** Proportion distribution of five levels of GG-RP in the primary cohort (PC), validation cohort (VC), testing cohort 1 (TC1), and testing cohort 2 (TC2). **(E-H)** Confusion matrix.

Compared with GG-NB, GG-Pre reduced the overall upgrading rate by 27.9% ( $P < 0.001$ ) and reduced the overall downgrading rate by 6.4% ( $P = 0.029$ ). The risks of upgrading and downgrading were diminished (**Figure 4B**). The consistency in GG 1 increased to 90.5% by applying our method ( $P < 0.001$ ). Furthermore, the proportion of GG 1 shifting to GG 5 was eliminated (**Figure 5A**). Compared with GG-NB, the upgrading rate in GG 2 decreased by 15% in our prediction ( $P = 0.093$ ). Moreover, shifts from GG 2 to 4 or 5 were significantly reduced. In GG 3, the upgrading rate dropped from 53.1% to 4.3% ( $P < 0.001$ ), and the downgrading rate decreased from 16.3% to 6.4% ( $P = 0.201$ ) (**Figure 5C**). The up-/downgrading rates of GG 4 were reduced by 18.5% ( $P = 0.112$ ) and 32.6% ( $P = 0.002$ ), respectively, and the proportion of shifting from GG 4 to 5 decreased by one third ( $P = 0.012$ ) (**Figure 5D**). Among all cases, the rates of GG raising two or more levels were reduced from 18.5% to 4.5%. Shifts from below GG 3 to 4 or 5 were reduced by 48.1%, among which the ratio of GG 1 shifting to 5 was eliminated. Among all grades, GG 3 was the group that obtained the highest cumulative gains for both upgrading and downgrading improvements.

### Discussion and Conclusion

For the past six decades, GS has remained one of the most powerful predictive factors for biochemical relapse and overall survival of PCa. Current treatment options are mainly decided via risk stratifications or nomograms, which consist of total PSA level, clinical T stage, NB Gleason Gs, and other clinico-pathological parameters. Thus, precise assignments of biopsy Gs are crucial when making optimal treatment choices for patients. However,

discrepancies between NB and RP pathology are common, and the latter is considered to reflect more accurate information about the nature of the tumor. Upgrading from NB to RP was reported to be as high as 36% [27], whereas downgrading was reported to have a lower average of 5% [1,8]. Tumor aggressiveness was usually underestimated in NB-upgraded cases, followed by worse prognoses of biochemical-free survival. Corcoran et al. reported that 28.6% of upgrading cases correlated with a higher risk of biochemical recurrence [28]. Boorjian et al. constructed a multivariate model to predict biochemical recurrence following RP in a cohort of over 8,000 patients, and the NB results demonstrated minimal additional value as compared with RP Gleason results [29]. Similar situations were observed in a Korean population cohort, in which upgraded cases demonstrated worse biochemical-free survival and worse metastasis-free survival [30].

The reasons for discordance between NB and RP are variable, such as tumor heterogeneity, sampling bias on needle biopsies, erroneous interpretation on inadequate tissue, and different practices of GG assignments at the core- or patient-levels. To achieve more accurate results of NB pathology, numerous attempts have been made. Several studies have tried to incorporate multiple clinical parameters (e.g., PSA, core length, percentage of Gleason pattern 4) to develop models or nomograms to predict final RP results. However, the robustness and discriminative power of these models remained below the desired threshold of 0.70 [31-33]. This situation was improved with the adoption of MRI in targeted biopsies (TBx). Level-1 evidence from the PRECISION trial demonstrated that MR fusion TBx improved the detection rate of clinically significant PCa [34]. TBx

can reduce upgrading and improve tumoral percentage at each core, compared with SBx [35-38]. Additionally, there is an increasing trend regarding the application of data science for automated GG scoring. Several automated deep learning algorithms for GG have been proposed on biopsy histology or tissue microarrays, producing accuracy ranges from 80% to 98% [39-41]. However, most of these studies focused on biopsy samples, and limited works have addressed upgrading NBs. Furthermore, to the extent of our knowledge, few works have been accomplished for predicting the final GG from an MRI to elucidate better discrepancies between NB and RP pathology [42-45].

The black-box feature of deep learning is often regarded as the main drawback of these artificial systems, especially for treatment-related decision-making in clinical practice. The method of voxel analysis has generally not been recommended for modeling the prostate MRI because of the weakened anatomical consistency caused by the distance between imaging layers and computational costs. Overwhelmed by redundant information, pixel-wise analysis of slices has also been a big challenge, owing to the naive statistics of slice-level results. To make our algorithm more interpretable and accurate, we proposed a radiologist-like computing framework for MRI for end-to-end prediction of GG-RP, named PCa-GGNet. This tool combined the dual advantages of the pixel-wise analysis of deep learning [46] and the dynamic programming of DRL [47]. In the current design, we defined each patient's T2WI-FS as a game, each imaging layer as a frame and each action as a gamer's movement. The current framework was only designed for a single sequence (T2WI-FS) in MRI, in which Diffusion-weighted imaging (DWI) or apparent diffusion coefficient (ADC) was not involved. The pre-process of center cropping in the original image expanded the proportion of the prostate in the input so that the model paid more attention to the prostate area. Quantitative and robust features combined with artificial intelligence, helped the framework draw a path for decision-making more quickly and accurately. First, to generate a patient-level result, the model's decisions were based not only on a specific single layer but also on the entire "impression" of imaging data at every previous step. This is how the human brain works. Second, the construction of the final decision consisted of both tumor volume and histological aggressiveness information to improve the discriminative power for the final GG (**Supplementary Information II**). For example, when there were two isolated lesions within one case at different layers (minor: 4 + 4; major: 3 + 3), if we choose the slice containing maximum tumor volume

as the decision basis, the patient-level decision would be 3 + 3. Otherwise, if we select the layer having the highest score, it would be 4 + 4 (**Figure S3**). Obviously, neither of the aforementioned two answers can be called accurate. However, by adding both tumor volume and histological ranking into the formula, our model successfully optimizes the recipe to mimic patient-level results (3 + 4), which is precisely the way radiologists do it.

Compared with radiomics [48]-based machine learning, PCa-GGNet better reflects the characteristics of RP pathology and avoids the restrictions of tumor segmentation. It improved the fitting quality of weakly supervised models and further reduced the dependence of annotated data. The primary principle of our design required the use of high-information entropy input modeling to compensate for the excessive reliance on supervisory information. This system should function well in scenarios that lack domain knowledge, especially for non-prominent tumors such as PCa. The transfer learning method [26] provided a powerful tool to artificial intelligence-based models for expanding advantages of the algorithm to different image types and have been proven in many clinical applications. The prediction model based on mp-MRI for multi-sequence (DWI, ADC, etc.) information joint was hopeful to be further constructed. The framework also has the potential to be extended to other medical image analysis tasks based on different modality images.

On our total dataset, the consistency rate of NB to RP was only 44.9%. Approximately 40.4% of NB cases upgraded to RP, and 14.7% finally downgraded. Patients falling in GG 1 are usually considered for active surveillance (AS) [49]. With the increasing percentage of Gleason pattern 4, patients are more likely to be referred to RP and other definitive therapies. Thus, upgrading from GG grade 1 to 2/3 is crucial to the selection between AS and definitive treatment. In our testing cohort, 64.6% of patients in GG 1 experienced upgrading at RP, which indicates that there is a possibility for a group having the same biopsy conditions to face insufficient treatment if they are recommended for AS. In contrast, when considering the implication of extended pelvic lymph node dissection (ePLND), upgrading from GG 2/3 to 4/5 is of great significance. In our testing cohort, over 23% of GG 2 and 53% of GG 3 upgraded to 4/5, meaning that the prognosis of these patients might be compromised, owing to the lack of ePLND. On the contrary, 14% of patients in GG 5 and 35.7% in GG 4 downgraded to GG 3 or below, which suggested that these patients might not benefit from ePLND during radical prostatectomy. Our developed algorithm significantly reduced both upgrading and down-

grading for every group. Our model obtained an ACC of approximately 0.85 with internal and external validation. Additionally, due to extended indication of RP in our institutions, we were lucky to have more proportions of lower Gleason patients in our cohorts (15% GG 1), which let our network more adaptive with low-grade GG cases. The critical function of our network was to find out and testify the internal correlations between imaging and pathological appearance. Thus, PCa-GGNet was also potential to be extended to a larger population with multiple objectives such as benign and malignant discrimination, significant and insignificant PCa distinguish tasks, and so on.

To further evaluate the additional benefits in real clinical practice, we stratified all patients in the validation cohort by total PSA, clinical T stage, and GG-NB/GG-Pre and constructed confusion metrics based on the 3-tier protocols of the National Comprehensive Cancer Network (NCCN) (**Supplementary Information III, Figure S4**). When compared with the GG-Pre-based stratifications, 28.4% (23/81) of the medium-risk patients in the GG-NB model would require ePLND for oncological control, and 4.4% (8/180) of the high-risk patients in the GG-NB model might not benefit from ePLND. The distribution differences between groups in the NCCN model were not as high as those in the GG metrics (**Figure 6**), which have the ability to reduce deviations derived from any individual parameter to achieve better performance for the whole system. Nevertheless, by improving the discriminative accuracy GG grading in the current study, we can achieve much better performance of the entire system (approaching the GG-RP model).

Several limitations must be mitigated. First, our current version only involved T2WI-FS data because of image standards and data scales. Future work should include more sequences (e.g., non-fs T2WI, DWI, ADC, and DCE) to provide better multi-tier GG prediction by transfer learning [26] and verify the proposed method on more international public data sets for different clinical applications (e.g., ProstateX dataset [50]). Second, multi-center validations at a larger scale of populations and prospective data were considered in the future. Third, the current study only involved systemic biopsy results. Thus, more work needs to be done to explore the discriminative power between our method and the targeted biopsy. Additionally, of all biopsy proved prostate cancer patients, only those who received radical prostatectomy had been enrolled in this study, which brought selection bias neglecting patients without RP surgery. Last, the decision-making process from core-level, specimen-level, to patient-level results is

variable. Some doctors tend to adopt the highest core-/specimen-level result for risk evaluation, whereas others tend to select the GG of the index lesion or provide a tertiary clinical decision. Ideally, it might be the best case to report precise percentages of each Gleason pattern for every lesion of all specimens. However, such work is nearly impossible for human efforts in clinical routines. During the deep learning era, further work on the quantitative Gleason pattern [51,52] evaluation at gland- or pixel levels should be explored.

In summary, we proposed a human-like PCa-GGNet framework to predict patients' final GG of RP. According to the multi-center validation, our method demonstrated high reliability of reducing risks of upgrading and downgrading from biopsy to RP pathology. This framework will facilitate clinicians by providing more precise treatment options, and it has the potential for application to other MRI-based tumor research.

## Abbreviations

ACC: five-grade prediction accuracy; CI: confidence interval; CNN: convolutional neural network; DRL: deep reinforcement learning; GS: Gleason score; GG: grade group; GG-NB: GG of needle biopsy; GG-RP: GG of radical prostatectomy; mp MRI: multiparametric magnetic resonance imaging; NB: needle biopsy; PC: primary cohort; RP: radical prostatectomy; T2WI: T2-weighted image; TC: testing cohort; VC: validation cohort.

## Supplementary Material

Supplementary figures and tables.

<http://www.thno.org/v10p10200s1.pdf>

## Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (No. 81922040, 61871004, 81930053, 81227901, 81527805, 31571001, 61828101), the National Key Research and Development Program of China (No. 2018YFC0115900), the Beijing Natural Science Foundation (No. 7182109), the National Key R&D Program of China (No. 2017YFA0205200), the Chinese Academy of Sciences (No. XDB32030200, XDB01030200, QYZDJ-SSW-JSC005), the Youth Innovation Promotion Association CAS (No. 2019136), Key Research and Development Project of Jiangsu Province (BE2018749), and Southeast University-Nanjing Medical University Cooperative Research Project (2242019K3DN08).

The authors thank the Department of Pathology and Department of Radiology, Peking University Third Hospital, and the Department of Urology,



Peking University People's Hospital, for their support with data collection.

## Author contributions

**Conceptualization:** Lizhi Shao and Ye Yan; **Data curation:** Lizhi Shao and Ye Yan; **Formal analysis:** Lizhi Shao and Ye Yan; **Funding acquisition:** Jie Tian; **Methodology:** Lizhi Shao, Guanyu Yang, and Zhenyu Liu; **Project administration:** Guanyu Yang, Zhenyu Liu, Yi Huang, Jian Lu, and Jie Tian; **Software:** Lizhi Shao; **Supervision:** Guanyu Yang, Lulin Ma, Zhenyu Liu, Xiongjun Ye, Jian Lu, and Jie Tian; **Validation:** Lizhi Shao, Yan Ye, Zhenyu Liu, Xuezhong Zhou, and Kai Sun; **Visualization:** Lizhi Shao; **Data collection:** Xiongjun Ye, Haizhui Xia, Xuehua Zhu, Yuting Zhang, Zhiying Zhang, Cheng Liu, and Wei He; **Data annotations:** Wei He, Min Lu, Huiying Chen, and Ye Yan; **Writing – original draft:** Lizhi Shao, Ye Yan, and Xiongjun Ye; **Writing – review & editing:** Lizhi Shao, Zhenyu Liu, and Guanyu Yang.

## Code availability

The source code of the proposed method is available from the Github repository: <https://github.com/StandWisdom/PCa-GGNet>.

## Competing Interests

The authors have declared that no competing interest exists.

## References

- Epstein JI, Feng Z, Trock BJ, Pierorazio PM. Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified Gleason grading system and factoring in tertiary grades. *Eur Urol*. 2012; 61: 1019-1024.
- Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. 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: 244-252.
- Loeb S, Folkvaljon Y, Robinson D, Lissbrant IF, Egevad L, Stattin P. Evaluation of the 2015 Gleason Grade Groups in a Nationwide Population-based Cohort. *Eur Urol*. 2016; 69: 1135-1141.
- Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. *Eur Urol*. 2016; 69: 428-435.
- He J, Albertsen PC, Moore D, Rotter D, Demissie K, Lu-Yao G. Validation of a Contemporary Five-tiered Gleason Grade Grouping Using Population-based Data. *Eur Urol*. 2017; 71: 760-763.
- Evans SM, Patabendi BV, Kronborg C, Earnest A, Millar J, Clouston D. Gleason group concordance between biopsy and radical prostatectomy specimens: A cohort study from Prostate Cancer Outcome Registry-Victoria. *Prostate Int*. 2016; 4: 145-151.
- Muntener M, Epstein JI, Hernandez DJ, Gonzalgo ML, Mangold L, Humphreys E, et al. Prognostic significance of Gleason score discrepancies between needle biopsy and radical prostatectomy. *Eur Urol*. 2008; 53: 767-775.
- Rajinikanth A, Manoharan M, Soloway CT, Civantos FJ, Soloway MS. Trends in Gleason score: concordance between biopsy and prostatectomy over 15 years. *Urology*. 2008; 72: 177-182.
- Gandaglia G, Ploussard G, Valerio M, Mattei A, Fiori C, Roumiguie M, et al. The Key Combined Value of Multiparametric Magnetic Resonance Imaging, and Magnetic Resonance Imaging-targeted and Concomitant Systematic Biopsies for the Prediction of Adverse Pathological Features in Prostate Cancer Patients Undergoing Radical Prostatectomy. *Eur Urol*. 2020; 77: 733-741.
- Vos EK, Litjens GJ, Kobus T, Hambroek T, Hulsbergen-van DKC, Barentsz JO, et al. Assessment of prostate cancer aggressiveness using dynamic contrast-enhanced magnetic resonance imaging at 3 T. *Eur Urol*. 2013; 64: 448-455.
- Mehralivand S, Shih JH, Rais-Bahrami S, Oto A, Bednarova S, Nix JW, et al. A Magnetic Resonance Imaging-Based Prediction Model for Prostate Biopsy Risk Stratification. *Jama Oncol*. 2018; 4: 678-685.
- Wang Q, Li H, Yan X, Wu CJ, Liu XS, Shi HB, et al. Histogram analysis of diffusion kurtosis magnetic resonance imaging in differentiation of pathologic Gleason grade of prostate cancer. *Urol Oncol*. 2015; 33: 315-337.
- Donati OF, Afaq A, Vargas HA, Mazaheri Y, Zheng J, Moskowitz CS, et al. Prostate MRI: evaluating tumor volume and apparent diffusion coefficient as surrogate biomarkers for predicting tumor Gleason score. *Clin Cancer Res*. 2014; 20: 3705-3711.
- Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, et al. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. *Radiology*. 2019; 293: 607-617.
- Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. *Theranostics*. 2019; 9: 1303-1322.
- Chaddad A, Kucharczyk MJ, Niazi T. Multimodal Radiomic Features for the Predicting Gleason Score of Prostate Cancer. *Cancers*. 2018; 10: 249-250.
- Nketiah G, Elschof M, Kim E, Teruel JR, Scheenen TW, Bathen TF, et al. T2-weighted MRI-derived textural features reflect prostate cancer aggressiveness: preliminary results. *Eur Radiol*. 2017; 27: 3050-3059.
- Jia P, Xue H, Liu S, Wang H, Yang L, Hesketh T, et al. Opportunities and challenges of using big data for global health. *Sci Bull*. 2019; 64: 1652-1654.
- Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck KSV, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019; 25: 1301-1309.
- Lucas M, Jansen I, Savci-Heijink CD, Meijer SL, de Boer OJ, van Leeuwen TG, et al. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch*. 2019; 475: 77-83.
- Cao R, Mohammadian BA, Afshari MS, Shakeri S, Zhong X, Enzmann D, et al. Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet. *IEEE Trans Med Imaging*. 2019; 38: 2496-2506.
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature*. 2015; 518: 529-533.
- Mahmud M, Kaiser MS, Hussain A, Vassanelli S. Applications of Deep Learning and Reinforcement Learning to Biological Data. *Ieee T Neur Net Lear*. 2018; 29: 2063-2079.
- Higa S, Iwashita Y, Otsu K, Ono M, Lamarre O, Didier A, et al. Vision-Based Estimation of Driving Energy for Planetary Rovers Using Deep Learning and Terramechanics. *IEEE ROBOTICS AND AUTOMATION LETTERS*. 2019; 4: 3876-3883.
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vision*. 2015; 115: 211-252.
- Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging*. 2016; 35: 1285-1298.
- Cohen MS, Hanley RS, Kurteva T, Ruthazer R, Silverman ML, Sorcini A, et al. Comparing the Gleason Prostate Biopsy and Gleason Prostatectomy Grading System: The Lahey Clinic Medical Center Experience and an International Meta-Analysis. *Eur Urol*. 2008; 54: 371-381.
- Corcoran NM, Hong MK, Casey RG, Hurtado-Coll A, Peters J, Harewood L, et al. Upgrade in Gleason score between prostate biopsies and pathology following radical prostatectomy significantly impacts upon the risk of biochemical recurrence. *Bju Int*. 2011; 108: 202-210.
- Boorjian SA, Karnes RJ, Crispen PL, Rangel LJ, Bergstralh EJ, Sebo TJ, et al. The impact of discordance between biopsy and pathological Gleason scores on survival after radical prostatectomy. *J Urol*. 2009; 181: 95-104.
- Bakavicius A, Drevinskaite M, Daniunaite K, Barisiene M, Jarmalaite S, Jankevicius F. The Impact of Prostate Cancer Upgrading and Upstaging on Biochemical Recurrence and Cancer-Specific Survival. *Medicina*. 2020; 56: 61-62.
- Chun FK, Steuber T, Erbersdobler A, Currlin E, Walz J, Schlomm T, et al. Development and internal validation of a nomogram predicting the probability of prostate cancer Gleason sum upgrading between biopsy and radical prostatectomy pathology. *Eur Urol*. 2006; 49: 820-826.
- Corcoran NM, Hovens CM, Hong MK, Pedersen J, Casey RG, Connolly S, et al. Underestimation of Gleason score at prostate biopsy reflects sampling error in lower volume tumours. *Bju Int*. 2010; 109: 660-664.
- Thomas C, Pfirrmann K, Piefles F, Bogumil A, Gillitzer R, Wiesner C, et al. Predictors for clinically relevant Gleason score upgrade in patients undergoing radical prostatectomy. *Bju Int*. 2012; 109: 214-223.
- Kasivisvanathan V, Emberton M, Moore CM. MRI-Targeted Biopsy for Prostate-Cancer Diagnosis. *N Engl J Med*. 2018; 379: 589-590.
- Goel S, Shoag JE, Gross MD, Al HAAB, Robinson B, Khani F, et al. Concordance Between Biopsy and Radical Prostatectomy Pathology in the Era of Targeted Biopsy: A Systematic Review and Meta-analysis. *Eur Urol Oncol*. 2020; 3: 10-20.
- Quentin M, Blondin D, Arsov C, Schimmoller L, Hiester A, Godehardt E, et al. Prospective evaluation of magnetic resonance imaging guided in-bore prostate biopsy versus systematic transrectal ultrasound guided prostate biopsy in biopsy naive men with elevated prostate specific antigen. *J Urol*. 2014; 192: 1374-1379.



37. Valerio M, Donaldson I, Emberton M, Ehdai B, Hadaschik BA, Marks LS, et al. Detection of Clinically Significant Prostate Cancer Using Magnetic Resonance Imaging-Ultrasound Fusion Targeted Biopsy: A Systematic Review. *Eur Urol.* 2015; 68: 8-19.
38. Elkhoury FF, Felker ER, Kwan L, Sisk AE, Delfin M, Natarajan S, et al. Comparison of Targeted vs Systematic Prostate Biopsy in Men Who Are Biopsy Naive: The Prospective Assessment of Image Registration in the Diagnosis of Prostate Cancer (PAIREDCAP) Study. *Jama Surg.* 2019; 154: 811-818.
39. Arvaniti E, Fricker KS, Moret M, Rupp N, Hermanns T, Fankhauser C, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep.* 2018; 8: 1-11.
40. Lucas M, Jansen I, Savci-Heijink CD, Meijer SL, de Boer OJ, van Leeuwen TG, et al. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch.* 2019; 475: 77-83.
41. Nguyen TH, Sridharan S, Macias V, Kajdacsy-Balla A, Melamed J, Do MN, et al. Automatic Gleason grading of prostate cancer using quantitative phase imaging and machine learning. *J Biomed Opt.* 2017; 22: 36015.
42. Antonelli M, Johnston EW, Dikaos N, Cheung KK, Sidhu HS, Appayya MB, et al. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. *Eur Radiol.* 2019; 29: 4754-64.
43. Abraham B, Nair MS. Computer-aided classification of prostate cancer grade groups from MRI images using texture features and stacked sparse autoencoder. *Comput Med Imaging Graph.* 2018; 69: 60-68.
44. Citak-Er F, Vural M, Acar O, Esen T, Onay A, Ozturk-Isik E. Final Gleason score prediction using discriminant analysis and support vector machine based on preoperative multiparametric MR imaging of prostate cancer at 3T. *Biomed Res Int.* 2014; 2014: 690787.
45. Fehr D, Veeraraghavan H, Wibmer A, Gondo T, Matsumoto K, Vargas HA, et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proc Natl Acad Sci U S A.* 2015; 112: 6265- 6273.
46. Litjens G, Kooi T, Bejnordi BE, Setio A, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017; 42: 60-88.
47. Petter EA, Gershman SJ, Meck WH. Integrating Models of Interval Timing and Reinforcement Learning. *Trends Cogn Sci.* 2018; 22: 911-922.
48. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by non-invasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5: 1-9.
49. Chun FK, Haese A, Ahyai SA, Walz J, Suardi N, Capitanio U, et al. Critical assessment of tools to predict clinically insignificant prostate cancer at radical prostatectomy in contemporary men. *Cancer-Am Cancer Soc.* 2008; 113: 701-709.
50. Armato SGR, Huisman H, Drukker K, Hadjiiski L, Kirby JS, Petrick N, et al. PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of medical imaging (Bellingham, Wash.).* 2018; 5: 4-5.
51. Sauter G, Steurer S, Clauditz TS, Krech T, Wittmer C, Lutz F, et al. Clinical utility of quantitative Gleason grading in prostate biopsies and prostatectomy specimens. *Eur Urol.* 2016; 69: 592-598.
52. Egevad L, Ahmad AS, Algaba F, Berney DM, Boccon Gibod L, Comp erat E, et al. Standardization of Gleason grading among 337 European pathologists. *Histopathology.* 2013; 62: 247-256.