Comprehensive AI Model Development for Gleason Grading: From Scanning, Cloud-based Annotation to Pathologist-AI Interaction

Xinmi Huo
Bioinformatics Institute, A*STAR, Singapore

KokHaur ONG
Bioinformatics Institute

Kah Weng Lau
Department of Pathology, National University Hospital, National University Health System, Singapore

Laurent Gole
Institute of Molecule and Cell Biology, A*STAR, Singapore

David Young
Institute of Molecule and Cell Biology, A*STAR, Singapore

Char Loo Tan
Department of Pathology, National University Hospital, National University Health System, Singapore

Chongchong Zhang
Department of Pathology, The 910 Hospital of PLA, QuanZhou, China

Yonghui Zhang
Department of Pathology, The 910 Hospital of PLA, QuanZhou, China

Xiaohui Zhu
Department of Pathology, Nanfang Hospital and Basic Medical College, Southern Medical University, Guangzhou, Guangdong Province, People's Republic of China.

Longjie Li
Bioinformatics Institute, A*STAR, Singapore

Hao Han
https://orcid.org/0000-0002-1551-2995

Haoda Lu
Bioinformatics Institute, A*STAR, Singapore

Gabriel Marini
Bioinformatics Institute, A*STAR, Singapore

Jun Xu
Nanjing University of Information Science and Technology

Kun Gui

Bingxian Chen

Wenzhao Shi

Wanyuan Chen

Zhejiang Provincial People's Hospital People's Hospital of Hangzhou Medical College

Stephan Sanders

University of California San Francisco https://orcid.org/0000-0001-9112-5148

Hwee Kuan Lee

Bioinformatics Institute, A*STAR, Singapore

Susan Swee-Shan Hue

Department of Pathology, National University Hospital, Singapore

Weimiao Yu (✉ yu_weimiao@bii.a-star.edu.sg)

Bioinformatics Institute, A*STAR

Soo Yong Tan

National University of Singapore

Article

Keywords: Digital Pathology, Artificial Intelligence, Deep Learning, Prostate Cancer, Pathologist-AI Interaction, Incremental Learning, Cloud Annotation, WSI quality check

Posted Date: September 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1917695/v1

Abstract

AI-based solutions for automated Gleason grading have been developed to assist pathologists to make rapid and quantitative assessments, but the generalization across various scanners and updating AI models continuously using new annotated data from end users remains a key bottleneck in the field. We proposed an comprehensive digital pathology workflow for AI-assisted Gleason grading, incorporating an image quality check software A!magQC, a cloud-based annotation platform A!HistoNotes and Pathologist-AI Interaction (PAI) strategy. To demonstrate and validate the pipeline, we employed it on prostate samples obtained from 5 scanners for Gleason grading. After training on 132 prostatectomy specimens scanned by an Akoya Biosciences scanner, validation on 55 prostatectomy specimens and 156 biopsy specimens yielded a sensitivity of 85%, specificity of 96% and F1 score of 78% on Gleason grading for prostatectomy specimens, and 96% sensitivity on tumor detection for biopsy specimens. For images scanned by other 4 scanners, the average F1 score increased from 67% to 75% on Gleason pattern detection after adopting our generalization solution. In clinical experiments conducted with 5 pathologists from Singapore and China, our pipeline accelerated Gleason scoring by 43%. Furthermore, it reduced annotation time by 60% via semi-automatic annotation, leading to improved model performance through incremental learning.

Introduction

Prostate cancer (PCa) is the second most common cancer of males worldwide, accounting for 15.1% of all male cancers diagnosed in 2020 [1]. The Gleason grade is the well-established histological parameter used to assess the aggressiveness of PCa. Accurate Gleason grading is crucial to establish a patient’s prognosis and determine the appropriate treatment. Large numbers of tissue samples generated by prostatectomy and routine core biopsy need to be carefully inspected under the microscope by experienced pathologists to detect malignancies. When PCa is detected, pathologists grade the tumor based on microscopic identification of histological characteristics, which is highly dependent on the skill, training and experience of the operator. To assess tumour load, the pathologist needs to measure the relative proportion of tumour by Gleason grade. This process is time-consuming and labor-intensive, especially when the number of samples is large, as with saturation prostatic biopsies. Misdiagnoses can result in preventable morbidity and mortality, whilst over-diagnosis can cause patient anxiety and lead to unnecessary procedures, including surgery. A better solution that leverages on technology is required to improve the efficiency, accuracy, and consistency of Gleason grading.

Recent advances in histology scanning technology and Artificial Intelligence (AI) offer great opportunities to improve the fidelity of Gleason grading. As shown in Figure 1A, the advent of different slide scanners has enabled the rise of digital pathology (DP) and the tantalizing promise of computational assessment. The use of Whole-Slide Image (WSI) has significantly expanded the volume of DP data by enabling digitalization of whole slides at high resolution. Meanwhile, AI holds the potential to improve the efficiency of pathological assessment by reducing turnaround time and enhancing detection consistency [2]-[5], as shown in the DP workflow in Figure 1A and 1B. Several traditional Machine Learning (ML) algorithms were investigated in gland segmentation and Gleason score by extracting morphological, textual, color, and other features to identify PCa within WSIs [6]-[10].

To fully utilize such massive data sets, deep learning (DL) has been adopted and shown to have superior performance in automatic feature extraction and pattern recognition, leading to widespread applications in detection, classification, and segmentation of tissues in WSIs [3]-[5]. Compared to traditional ML approaches, DL based methods can learn more complex Gleason patterns from histopathological images. Recent research focus on applying DL to alleviate pathologists' workloads, improve diagnostic accuracy, and reduce observer variations have shown promise [11]-[25]. For example, Paige Prostate Alpha is a PCa detection system based on a weakly-supervised DL algorithm. Independent studies on it focused on comparing the accuracy and efficiency of PCa detection with and without AI assistance [20]-[25] have shown that the sensitivity for pathologists increased from 74% to 90%, with reduction of diagnostic time by 65.5%. It is worth noting that Paige Prostate Alpha is classified as a Class II device by the U.S. Food and Drug Administration (FDA), the first FDA-approved AI product in the field of DP.

Despite substantial achievements, we still face challenges with development and deployment of AI technology in histopathology practice. First, the variation and quality of images acquired from different scanners may profoundly affect AI models [26], which is an unrecognized key variable that impacts the first step in DP pipelines as shown in Figure 1A. Some quality issues, such as out-of-focus and artifacts, may impair the database and bias the resulting AI-model. Currently, we lack a standard by which to select scanned images for AI training. Furthermore, the impact of scanners on WSI analysis and AI models has not been systematically studied. Most previous studies [11]-[25] only assessed the performance of AI models on images scanned by one or two scanners. For example, Paige Prostate Alpha was validated on images scanned by Leica and Philips only. Second, large quantities of annotated data are essential for training DL models [5], but acquiring such annotations is expensive. It requires experienced pathologists to take time from their heavy clinical workload to manually perform detailed annotations. Last, the feasibility of updating existing AI-based methods with more data should be investigated since the performance of models can be improved by increasing the amount of training data without completely retraining the models, thereby improving their cost-effectiveness and reducing validation time [27][28].

To meet these challenges, we have developed and validated a pipeline as conceptually illustrated in Figure 1A-C. Given the expense of generating a large, highly curated and annotated database, it should be developed based on the highest quality WSIs available. As shown in Figure 1B, the A!MagicQC quality control software package was developed for DP image quality control to ensure that the Hematoxylin and Eosin (H&E) WSIs used in this study are of high quality (details of A!MagicQC described in Supplementary Section 1). We also quantified the image quality characteristics of several popular scanners. To address the need for an integrated and user-friendly annotation platform, we developed A!HistoNotes, a cloud-based tool that allows pathologists worldwide to perform annotations or view the predictions generated by the AI model in a web browser. To facilitate rapid incorporation of new data and incremental learning, we closely integrated A!HistoNotes with our AI model in a feedback loop we have termed Pathologists-AI Interaction (PAI) [29][30]. This closed loop enabled semi-automatic annotation to ease the annotation burden and facilitate AI model development, as demonstrated in Figure 1C. After training, the AI model was first validated on a prostatectomy and biopsy dataset scanned by Akoya Sciences for tumor identification and grading. Subsequently, the model was extended to images scanned by multiple scanners through normalization methods, which enabled consistent performance of our model across different scanners. Additional clinical validation, which included AI-assisted diagnosis and semi-automatic annotation followed by incremental learning, was conducted with pathologists from Singapore and China to evaluate the proposed method in a real-world application.

Materials And Methods

Sample preparation

Prostatectomy and biopsy formalin-fixed paraffin-embedded (FFPE) tissue specimens were collected from the Department of Pathology, National University of Singapore Hospital (NUHS) approved by the NHG Institutional Review Board (DSRB study reference number: 2018/01186). The samples were prepared in accordance with the standard operating procedures of a CAP-accredited histopathology laboratory. $4\mu m$ sections were cut and stained with hematoxylin and eosin from tissue blocks of radical prostatectomy and biopsy specimens. In total there were 187 prostatectomy specimens (tissue area 112,400mm^2) and 156 biopsy specimens (tissue area 7,723mm^2) from 214 patients. The profiles of patients involved in the study are listed in Supplementary Table 1.

WSI Scanning and Image Quality Control using A!magQC

Images were initially acquired using Vectra ® Polaris™ from Akoya Bioscicences with bright-field imaging at 0.5μm × 0.5μm per pixel resolution (20× magnification). Typical image size is approximately 85,000 × 40,000 pixels for prostatectomy specimens and 25,000 × 15,000 pixels for biopsy specimens. Since annotation quality and subsequent model development are ultimately dependent on the quality of scanned WSIs, it is critical to have a robust QA system to control for the variation of sample preparation and tissue types. Furthermore, image appearance and quality might significantly vary among different scanners, leading to differences in color, brightness, and contrast for the same samples. Different pathology laboratories may also use other brands of scanners, such as Olympus, Leica, KFBio, and Philips. To address these variables, we developed an image quality control software named A!magQC to identify common image quality issues for images acquired from various scanners. A!magQC is a fully automated histology image quality assessment tool to identify five common categories of WSI quality issues: out of focus, low contrast, saturation, artifacts, and texture uniformity, *etc.*[31]-[36]. Details of A!magQC are described in Supplementary Section 1. To study the characteristics of images scanned by different scanners and their impact on AI model performance, 38 prostatectomy specimens were randomly selected and scanned using 4 other scanners.

Structured Image Data Annotation using A!HistoNotes

With our image quality control tool, we ensured the scanned images for annotation are of high quality. A user-friendly, easily accessible, and efficient annotation solution is desired for structured and detailed annotation for AI model development and validation. To this end, we developed a cloud-based annotation platform, called A!HistoNotes. It provides a region of interest (ROI) management system and can create an adjustable ROI on top of the image viewer. Details of A!HistoNotes can be found in Supplementary Section 2.

After we uploaded WSIs to A!HistoNotes, three experienced pathologists from NUH in Singapore manually annotated the images independently to yield total annotation areas of 22,148 mm^2 (12,630 instances) on 187 prostatectomy specimens and 2,223 mm^2 (2,852 instances) on 156 biopsy specimens, as illustrated in Figure 2A-B. Class labels include Gleason pattern 3 (GP3), Gleason pattern 4 (GP4), Gleason pattern 5 (GP5), benign, and stroma tissue. For binary classification, we further grouped the GP3, GP4, and GP5 tissue as “Malignant,” whereas benign and stroma tissue comprise the “Non-malignant” group.

Patch Extraction and Preparation

The annotations of different classes were extracted from A!HistoNotes and organized, as illustrated in Figure 2C. Pathologists make diagnoses by examining specimens under a microscope at different magnifications, suggesting that the AI model might also have an optimal magnification for patch classification. To study and optimize patch scale factors, image patches were cropped into different sizes of overlapping patches using a sliding window approach, with side length = x and stride = (x-1)/2, where x = 251px (~125μm), 501px (~250μm), 1001px (~500μm), and 2001px (~1000μm), as illustrated in Figure 2D. The patches were then resized to 224 × 224 pixels to fit the input layer, resulting in a resolution of 0.56 (extra high resolution), 1.12 (high resolution), 2.24 (medium resolution), and 4.47 (low resolution) μm/pixel for patches whose original side length was 251, 501, 1001, and 2001 pixels, respectively. The overlap ratio between adjacent image patches is $(x+1)/2x$. To ensure validity, image patches were extracted only when more than 70% of the area was annotated.

Considering that prostatectomy specimens are much larger and thus contain more information than biopsies, we used the annotations in prostatectomy WSIs to train our models. The 187 radical prostatectomy WSIs were split into training and testing sets, whilst the annotated biopsy images were used for testing only. The training and testing set split ratio of prostatectomy WSIs is 7:3 for the number of WSIs, evenly divided to ensure the same ratios of areas for each annotated class in both training and testing since the annotated area of different classes may vary significantly from slide to slide. The process of train-test splitting is described in Supplementary Table 2. Figure 2E shows the number of patches of high resolution in the training and testing datasets for each class.

Scale optimization and AI model development

To determine the structure of the AI model, we first chose three of the most widely used network architectures in the field, namely, ResNet50, VGG16, and NasNet Mobile, to compare their performance at high resolution. Thereafter, the scale factor was optimized by training models built on the selected structures using images of the different resolutions from scratch. After the image patches for the training were organized as shown in Figure 3A, random data augmentation such as rotation and flipping were applied to each image patch prior to training as shown in Figure 3B. The classification layer in the network was replaced with a Weighted Classification Layer as a class rebalancing strategy, as demonstrated in Figure 3C, where the weights are inversely proportional to the number of image patches to mitigate imbalance in the dataset.

In the testing process, the trained model was applied to every sliding window on test images within the tissue region. The testing returned a list of scores that indicate the predicted probabilities of class labels for each patch, as demonstrated in Figure 3 D-E. We designed a voting policy to determine the predicted label of the overlapping region based on the predictions of neighboring patches. The final decision is either the class with highest votes (if there is any) or based on the highest average scores. Details of the voting policy are described in Supplementary Table 3. A final output was generated after the voting policy and compared to the ground truth annotation for performance evaluation, as shown in Figure 3F and G. Details of evaluation on testing images are described in Supplementary Section 3.

We performed AI model training and testing on MATLAB 2021a (Mathworks Inc., USA) with its Deep Learning and Image Processing toolboxes on the Windows 10 x64 operating system. The computer specifications are RAM: 1.0TB, CPU: Intel(R) Xeon(R) Gold 6242 CPU @ 2.80GHz, and GPU: single NVIDIA Tesla V100-PCIE-16GB.

Model Generalization Across Different Scanners

Since a variety of scanners are used in hospitals and laboratories with potential differences in resultant images, we investigated the generalizability of the optimal AI model on WSI images from 5 major commercially available scanners as shown in Figure 1A.

To normalize images across scanners and reduce the inconsistencies of image appearances, 132 patches of 2000 × 2000 pixels were randomly extracted from 66 Akoya-scanned WSIs in the training set as references. We transformed the images from other scanners by mapping the histogram of the references. Next, we enlarged the training data set using color augmentation as shown in Figure 3B to train a more generalizable model. By color argumentation, the resulting patches simulated the range of appearances acquired by different scanners. Three configurations are applied to each original image patch, and the details are described in Supplementary Table 4.

The images scanned by the 4 other scanners were not annotated, and the annotations on images scanned by the Akoya Biosciences scanner cannot be directly applied to them since the images were not aligned with each other. We applied image registration to migrate the annotations to the images of other scanners. After registration, images scanned by other scanners have corresponding ground truth annotations and, therefore, can be compared with the prediction results to evaluate the performance before and after model generalization. The average structural similarity index measure (SSIM) is 0.99 between original annotations and registered annotations, indicating accurate image registration.

Results

Scale optimization, tumor detection and Gleason grading

Architectures of different models were first trained using high resolution patches (1.12μm/ pixel). Comparing the model performances on image patch level (as shown in Supplementary Figure 3), NasNet Mobile is slightly inferior in terms of F1 score, whilst the differences between vgg16 and ResNet 50 are subtle. Given that the network size of ResNet50 is smaller, it was selected for faster deployment.

Using structures based on Resnet 50, models were trained at four different scales and applied to test images to compare their performance and optimize the scale factors. A representative image and the results of different scales are shown in Figure 4A. As demonstrated in Figure 4B, the resolution/scale will affect the performance of the AI model, and we found the scale factors have less impact on the detection of G3, G4, and non-Malignant classes. However, at high resolution, the F1 score of G5 increased more than 2-fold from low resolution as shown in Figure 4B. These findings correspond to the fact that most benign and malignant tissues can be easily identified at low magnification based on glandular architecture, whereas pathologists need higher magnification for determination of Gleason grading. We therefore selected the model trained using high-resolution patches (1.12μm/ pixel) to process prostatectomy specimens in the following study. The specificity of predicting GP3, GP4, Gp5 and non-malignant tissue is each greater than 0.9. Of note, the precision, sensitivity and F1 score of predicting non-malignant tissue is 0.99, which means the model rarely misclassified tumor as benign tissue. The training performance of high-resolution patches is demonstrated in Supplementary Figure 4.

Besides grading tumor on prostatectomy specimens, another key purpose of the AI model is to pre-screen the biopsy sample to highlight potential malignancies for pathologists before their review. For this screening, we applied the optimized model to test the performance of tumor detection (binary classification) on the biopsy specimens. Considering that the size of each biopsy is much smaller than that of prostatectomy specimen, and biopsy shapes are more elongated, biopsy specimens were processed at high and extra high resolution to determine if the model requires higher resolution to interpret biopsies. As shown in Figure 4C, results at extra high resolution are more detailed and consistent with ground truth annotation. As shown in Figure 4D, the differences of specificity and PPV between the two models are small, but the extra high resolution achieved significantly higher sensitivity and NPV. Therefore, the model using extra high-resolution images is more desired for biopsy specimens as the model can process the smaller tissue with irregular shape and achieved better performance. The precision of predicting benign tissues (NPV) was 0.96 for 156 biopsy samples, meaning that tumors are very unlikely to exist when the AI model has classified as non-malignant. This binary model can thus be applied as a pre-screening tool, confidently identifying malignancies from a large number of otherwise benign cases before manual screening by pathologists so that they can spend less time examining cases reported as benign and devote more attention to potential malignancies.

Model Generalization across different scanners

Images acquired from different scanners do have substantial variation, as demonstrated in Figure 5A. Using A!magQC to analyze image quality, we also found that images scanned by these scanners have varying quality issues, as shown in Figure 5B. Although images acquired from KFBio and Olympus scanners had more artifacts, the general quality of all images scanned by different scanners are satisfactory.

The generalizability of AI models across image scanners is a crucial consideration when deploying the AI model for diagnosis in hospitals using existing pathology workflows and scanners, ideally without the need to retrain the model. There is currently scarce data on the impact of different scanners on AI-enhanced digital pathology. Our data showed that uncompensated variations in the intensity and color of commonly used scanners will lead to inconsistent predictions, as shown in Figure 5C. In this study, normalization and color augmentation were applied to improve the generalizability of our model for different scanners. As demonstrated in Figure 5D, the appearance of image output by the existing AI model is more consistent after the generalization method had been applied. For quantification, the overall sensitivity, specificity, and F1 score for tumor grading significantly increased after generalization, as shown in Figure 5E-G. For images scanned by KFBio (III), and Leica (IV), the improvement is even more apparent, as the difference between these images and those scanned by the Akoya scanner is more dramatic. Although the training data was derived from only a single scanner, image normalization and color augmentation effectively improved the model consistency and performance across different scanners.

3-phase clinical validation of AI-assisted diagnosis

We conducted clinical validation in 3 phases involving 5 pathologists from China and Singapore to assess if the AI model can improve the accuracy, efficiency, and consistency of Gleason grading in histopathology departments, as shown in Figure 6A. Phase 1 was the conventional microscopic examination, while phase 2 was WSI examination without AI-assistance. In phase 3, annotations generated by the AI model, which we called pseudo annotations, as well as Gleason score and tumor percentages generated by the AI model are provided as references for pathologists. It should be noted that the displayed pseudo annotations have been simplified for easier interpretation and quicker diagnosis based on pathologists' feedback prior to the experiment, as shown in Figure 6B. Details of this experiment are described in Supplementary Section 4.

We first assessed the accuracy of Gleason grading by pathologists across the 3 phases, as indicated in Figure 6C. The average accuracy of pathologists in phase 1-3 is 0.67, 0.52 and 0.62, respectively, compared with an overall AI model accuracy of 0.68. With AI-assistance, the accuracy of three pathologists (B, D and E) vastly increased, while pathologists A remained highly accurate at all phases.

As for time taken for histological examination, almost no difference was observed between phase 1 and 2 among pathologist A, B and C, as shown in Figure 6D. Nevertheless, there was significant improvement for all 5 pathologists in phase 3, especially for pathologists A, D and E, whose average examination time was reduced by 41%-58%. The average examination time of pathologists decreased from 148s (phase 1) and 147s (phase 2) to 84s (phase 3), after AI-assistance was introduced in the final phase. These results show that WSI examination without AI-assistance is comparable to microscopic examination for Gleason grading as neither efficiency nor accuracy improved. However, WSI examination enhanced by the AI model accelerated Gleason grading time by 43% while maintaining the same high accuracy as with conventional microscopic examination, suggesting a role as physician extender to assist pathologists in their diagnostic work.

A common challenge for pathologists is inconsistency in classifying Gleason patterns among practitioners, as diagnosis may be influenced by the training, experience, and attentiveness of pathologists, leading to considerable inter-observer variability. As demonstrated in Figure 6E, compared to phase 1. the intraclass correlation coefficient (ICC) of three pathologists in phase 2 decreased slightly. However, comparing phases 2 and 3, where all 5 pathologists participated, the ICC increased from 0.78 to 0.84, indicating higher agreement among 5 pathologists when AI-assistance was provided based on improved accuracy and efficiency.

From the result of these experiments, we demonstrated that using WSI examination with AI-assistance significantly improved the accuracy, efficiency, and consistency of Gleason grading.

Pathologist-AI Interaction via A!HistoNotes: semi-automatic annotation and incremental learning

Repeated rounds of AI model training with new data progressively optimizes the model's performance. During this training, pathologists can leverage the existing model's classifications to generate pseudo annotations, which serve as a starting point for further refinement by pathologists. We have termed this iterative training Pathologist-AI Interaction (PAI) and implemented it in A!HistoNotes, where pathologists can correct pseudo annotations generated by the AI model and these annotations can be directly applied to the model for further training. Details of Pathologist-AI Interaction study are described in Supplementary Section 5.

To validate this PAI, we conducted a round of iterative training. First, we compared annotated areas before and after pathologists' corrections to quantify the quality of pseudo annotations generated by the AI model. Pathologists were asked to delete the instances they consider invalid. As shown in Figure 6F, most of the pseudo annotations of GP3, GP4, normal, and stroma were preserved, while many annotations of GP5 were removed by the pathologists. These findings were consistent with the image dataset's validation results, where the performance on identifying GP 3, GP 4 and non-malignant is satisfactory but relatively poor for GP5 due to the limited amount of GP5 data available. Regarding the time required for annotations by pathologists, the average annotation speed decreased from 1267s /per image for fully manual annotation, to 508s /per image for semi-automatic annotations (see Figure 6G), showing that semi-automatic annotations is about 60% faster than fully manual annotations. These findings are in line with the pathologists' preservation of most of the pseudo annotations generated by the models as the pathologists did not need to spend much time and effort to delete the annotations. From this perspective, semi-automatic annotations supported by our AI model greatly improve a pathologist's efficiency to iteratively integrate new data into the model.

More importantly, the overall performance improved after feeding the corrected annotations directly into the existing model for further training, as illustrated in Figure 6H. All metrics increased after the first round of PAI interactions although only a small set of data was added, demonstrating our model's capacity for incremental learning ability. This feature enables us to iteratively update the AI model without the need for complete retraining. As the performance of this AI model will be improved continuously by learning from pathologists' corrections, we also expect that pathologists will require substantially less time and effort to perform semi-automatic annotation and the accuracy of pseudo-annotations will gradually increase.

Conclusion And Discussion

Training multiple models with different resolutions performed well at identifying benign tissue, whereas ability to identify GP5 varied significantly. For GP5, we found that, the higher the image resolution yielded a more sensitive model but at the expense of specificity. The reason might be that GP 5 needs examination at high magnification to be identified, while benign tissue can be easily confirmed at lower magnification. Another possible reason explanation may be that, in our data set, the average size of GP5 annotations is much smaller than others, thereby requiring smaller image patches to process.

After selecting the optimal scale, we demonstrated that our AI model achieved high performance on tumor detection and grading on both biopsy and prostatectomy specimens. For prostatectomy specimens, our AI model provided pseudo annotations, tumor identification and percentage of various Gleason patterns as references for pathologists. In particular, the more abundant GP3 and GP4 samples afforded a specificity of more than 95%, which is clinically relevant because GP3 and GP4 are the most prevalent Gleason grades in most PCa in Singapore. For biopsy specimens, pathologists need to be extremely careful to identify even a small amount of tumor even though most cases are benign in practice. Our AI model can be used as a pre-screening tool to filter out suspicious cases and allow pathologists to review benign specimens quickly with greater confidence, which will reduce their workload and improve clinical turnaround times.

Of the Gleason grades, our model encountered the greatest difficulty with precise identification of GP5. GP5 is relatively rare, in keeping with it being the smallest group in our data, which adversely impacts our AI model's ability to differentiate it from other patterns. However, as more GP5 data is collected in future, it will enhance and validate this performance.

In addition, we have demonstrated that our model retains consistent and satisfactory performance across a range of major scanners, even though it had been trained using images scanned by Akoya Bioscience only. It is crucial that the model should be validated on images acquired from different scanners as the variation caused by image scanning can greatly affect the AI model performance unless a generalization approach is applied, as demonstrated in the results section (see Figure 5). The ideal AI model should be agnostic to the type or brand of scanner, allowing the DP solution to be implemented in different hospitals and labs without requiring replacement of existing equipment. In theory, a generalizable model would also be more tolerant of changes in imaging parameters that will likely occur in real-world usage, including suboptimal tissue processing and scanning conditions.

In the 3-phase experiment conducted with five pathologists from Singapore and China, we demonstrated that the AI model helped to increase the accuracy and efficiency of Gleason grading, whilst reducing inter-observer variation. This is particularly true in the efficiency analysis, where all five pathologists achieved faster times without a reduction in accuracy compared to microscopic examination and WSI examination without AI-assistance. Even though pathologists were familiar with microscopic examination and are new to WSI examination with AI-assistance, they still benefited greatly from the information provided by our AI model integrated with A!HistoNotes.

Results of the PAI experiment show that AI assistance significantly increases the speed of annotations by 60% compared to conventional manual annotation, which can lighten pathologists' workload. This feature is critical, since the training of AI models typically requires large amounts of annotated data, and researchers are unlikely to collect sufficient data to train a perfect model in a short period, especially given the cost of manual annotations. By enabling an iterative feedback loop between AI model and pathologists through the annotation platform provided by A!HistoNotes and incremental learning, the AI model can be further trained and validated in a more efficient way whilst building a larger, generalizable database.

In this study, all our annotations and clinical assessments were performed using A!HistoNotes. Pathologists used A!HistoNotes to performed manual annotation initially, and semi-automatic annotation after the first model was built. This cloud-based and user-friendly platform is easy for pathologist to access and annotate from different countries. In the 3-phase clinical validation study, the results show that the AI-assistance implemented in A!HistoNotes enables pathologists to locate and grade tumors, thereby greatly reducing the time required for pathologists' evaluation.

It is noteworthy that the format of AI-assistance was adjusted several times prior to this experiment according to pathologists' feedback. In the beginning, a more exhaustive list of pseudo annotations and Gleason score calculated by the model was presented to pathologists. However, they provided feedback that there was too much information to be interpreted in a timely manner, and that they were unlikely to benefit from such assistance as it was much more tedious and time-consuming than microscopic examination. After adjusting the pseudo annotations by applying morphological operations and adding more statistical measurements such as tumor and Gleason pattern percentage, we finally came up with a method to provide AI-assistance that pathologists considered useful. As this solution is meant to assist pathologists, it is crucial to design a user-friendly platform to connect the AI model with pathologists.

For the researchers' perspective, the ground truth annotation visualized in A!HistoNotes is an important reference to develop AI models as it allows them to quickly gain a brief understanding of many annotated structures. Therefore, A!HistoNotes may serve not only pathologists, but also researchers in the field of DP to develop and deploy AI solutions efficiently.

Future work will focus on integrating the inter-observer variations in annotation of Gleason Patterns into model training. Furthermore, it will be necessary to validate and optimize our model for a broader range of clinical scenarios, including varying tissue thickness, a wider variety of staining, and a broader array of scanners. Last but not least, the existing workflow of AI model development including A!magQC and A!HistoNotes can be applied to the development and validation of AI-based diagnostics for other types of cancer.

Declarations

Contributions

WY and TSY conceived the study, supervised the research together with SH, contributed to project coordination and administration. SH, WY and HH collected the data. XH contributed to model development and validation, experiment design, data analysis. SH and OKH contributed to the development and maintenance of A!HistoNotes and provided informatics support. LG contributed to the development of A!magQC, and together with LL provided informatics support. SH, LKW, TCL performed annotations and participated in the clinical experiments. CZ, YZ and XZ participated in the clinical experiments. XH wrote the manuscript with the contribution of all other authors. DY, HL, GM, XJ, KG, BC, WS, WC, SS, WY and TSY contributed to the review and revision of the manuscript.

Declaration of Interest

All authors declare no competing interests.

Data Sharing

The image data set will be made will be published for Automated Gleason Grading Challenge 2022 as a registered MICCAI 2022 challenges. The dataset can be downloaded at the challenge website:

https://aggc22.grand-challenge.org. The data can be used under an Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. Anyone who uses the data should cite the current article.

Acknowledgments

This project is funded by Exploit Technologies Gap-funded Project: “AI based H&E image

analysis for prostate cancer staging” (EPTL/18-Gap029-R20H) and A*STAR BMRC ATR Grant.

References

1. Ferlay, J. *et al.* Cancer statistics for the year 2020: An overview. *International journal of cancer* **149**, 778–789 (2021).
2. Di Cataldo, S. & Ficarra, E. Mining textural knowledge in biological images: Applications, methods and trends. *Computational and structural biotechnology journal* **15**, 56–67 (2017).
3. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology* **16**, 703–715 (2019).
4. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nature medicine* **27**, 775–784 (2021).
5. Alom, M. Z. *et al.* A state-of-the-art survey on deep learning theory and architectures. *Electronics* **8**, 292 (2019).
6. Nguyen, K., Jain, A. K. & Allen, R. L. Automated gland segmentation and classification for gleason grading of prostate tissue images. In *2010 20th International Conference on Pattern Recognition*, 1497–1500 (IEEE, 2010).
7. Naik, S., Doyle, S., Feldman, M., Tomaszewski, J. & Madabhushi, A. Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information. In *MIAAB workshop*, 1–8 (Citeseer, 2007).
8. Diamond, J., Anderson, N. H., Bartels, P. H., Montironi, R. & Hamilton, P. W. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human pathology* **35**, 1121–1131 (2004).
9. Farjam, R., Soltanian-Zadeh, H., Zoroofi, R. A. & Jafari-Khouzani, K. Tree-structured grading of pathological images of prostate. In *Medical Imaging 2005: Image Processing*, vol. 5747, 840–851 (International Society for Optics and Photonics, 2005).
10. del Toro, O. J. *et al.* Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. In *Medical Imaging 2017: Digital Pathology*, vol. 10140, 1014000 (International Society for Optics and Photonics, 2017).

11. Bulten, W. *et al.* Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology* **21**, 233–241 (2020).
12. Lucas, M. *et al.* Deep learning for automatic gleason pattern classification for grade group determination of prostate biopsies. *Virchows Archiv* **475**, 77–83 (2019).
13. Tolkach, Y., Dohmg¨orgen, T., Toma, M. & Kristiansen, G. High-accuracy prostate cancer pathology using deep learning. *Nature Machine Intelligence* **2**, 411–418 (2020).
14. Nagpal, K. *et al.* Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine* **2**, 1–10 (2019).
15. Nagpal, K. *et al.* Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens. *JAMA oncology* **6**, 1372–1380 (2020).
16. Pantanowitz, L. *et al.* An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health* **2**, e407–e416 (2020).
17. Arvaniti, E. *et al.* Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports* **8**, 1–11 (2018).
18. Bulten, W. *et al.* Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. *Modern Pathology* **34**, 660–671 (2021).
19. Singhal, N. *et al.* A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Scientific reports* **12**, 1–11 (2022).
20. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25**, 1301– 1309 (2019).
21. Raciti, P. *et al.* Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Modern Pathology* **33**, 2058–2066 (2020).
22. da Silva, L. M. *et al.* Independent real-world application of a clinical-grade automated prostate cancer detection system. *The Journal of pathology* **254**, 147–158 (2021).
23. Perincheri, S. *et al.* An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Modern Pathology* **34**, 1588–1595 (2021).
24. Dogdas, B. *et al.* Computational pathological identification of prostate cancer following neoadjuvant treatment. (2020).
25. Kanan, C. *et al.* Independent validation of paige prostate: Assessing clinical benefit of an artificial intelligence tool within a digital diagnostic pathology laboratory workflow. (2020).
26. Leo, P. *et al.* Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. *Journal of medical imaging* **3**, 047502 (2016).
27. Polikar, R., Upda, L., Upda, S. S. & Honavar, V. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)* **31**, 497– 508 (2001).

28. Joshi, P. & Kulkarni, P. Incremental learning: Areas and methods-a survey. *International Journal of Data Mining & Knowledge Management Process* **2**, 43 (2012).

29. Lutnick, B. *et al.* An integrated iterative annotation technique for easing neural network training in medical image analysis. *Nature machine intelligence* **1**, 112–119 (2019).

30. Holzinger, A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* **3**, 119–131 (2016).

31. Venkatanath, N., Praneeth, D., Bh, M. C., Channappayya, S. S. & Medasani, S. S. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, 1–6 (IEEE, 2015).

32. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. Histoqc: an open-source quality control tool for digital pathology slides. *JCO clinical cancer informatics* **3**, 1–7 (2019).

33. Ameisen, D. *et al.* Automatic image quality assessment in digital pathology: from idea to implementation. In *IWBBIO*, 148–157 (2014).

34. Bradley, D. & Roth, G. Adaptive thresholding using the integral image. *Journal of graphics tools* **12**, 13–21 (2007).

35. Pech-Pacheco, J. L., Crist´obal, G., Chamorro-Martinez, J. & Fern´andezValdivia, J. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 3, 314–317 (IEEE, 2000).

36. Haralick, R. M., Shanmugam, K. & Dinstein, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics* 610–621 (1973).

Figures

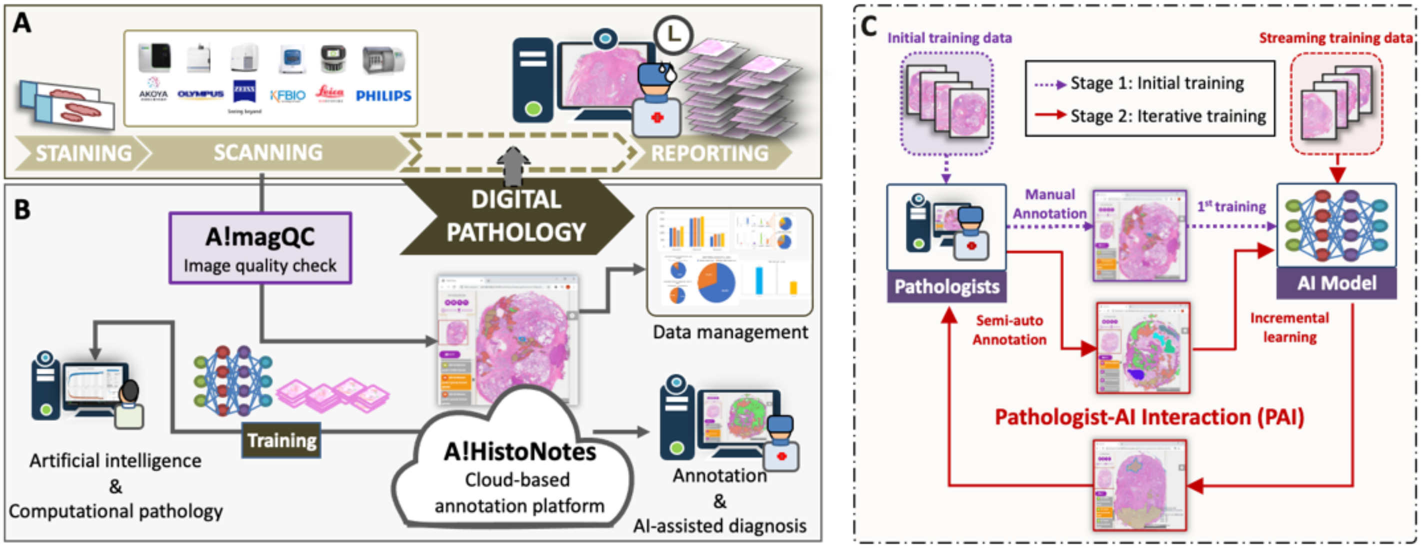

Figure 1 Overview The exiting digital pathology assessment pipeline is demonstrated in **(A)**. We designed an integrated and comprehensive Digital Pathology (DP) image analysis pipeline powered by A!magQC (image quality check), A!HistoNotes (DP image viewer, annotation platform, and database), and AI model (prostate tumor and grading identification), as shown in **(B)**. **(C)** is the workflow of pathologist-AI interaction (PAI). In this study, we implemented two stages of the PAI strategy: (1) Stage 1: pathologists perform manual annotation using A!HistoNotes, and a base model is trained using these annotated data. This process performed only once. (2) Stage 2: the base model is applied to data not used for the annotations and generates pseudo annotations on WSIs. Pathologists review and modify the imperfect annotations using A!HistoNotes. The corrected annotations are fed back to the model for further training. This is a repeated process until the model achieves high accuracy. After certain criteria is met, the model can generate accurate annotation and assist pathologists in clinical diagnosis.

Figure 1

See image above for figure legend.

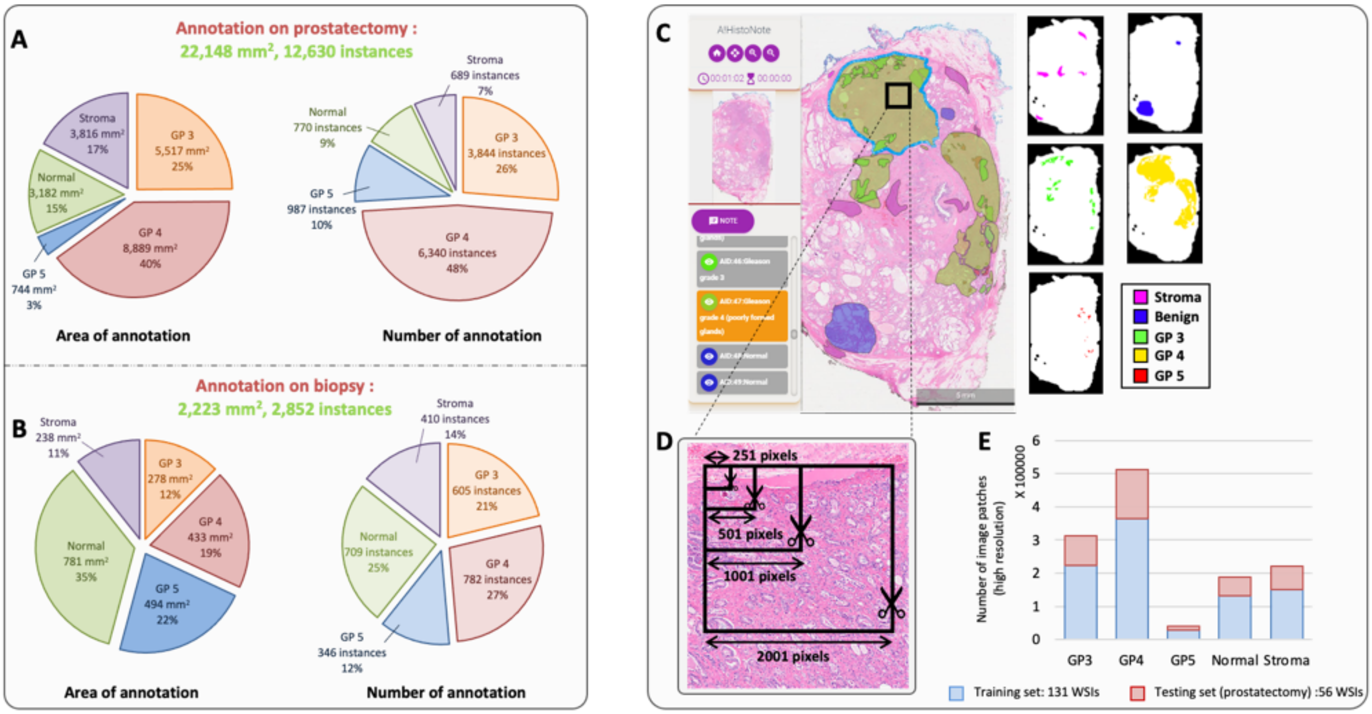

Figure 2 **Data profile** The amount of annotation performed on prostatectomy and biopsy by pathologists manually is summarized in **(A)** and **(B)**, respectively. Annotation of different classes was extracted from A!HistoNotes, and cropped into different sizes for scale optimization, as demonstrated in **(C)**. Prostatectomy specimens were split in to training and testing set, while biopsy specimens are used for testing only. Same configuration of train-test splitting was applied to dataset of different scales. **(D)** shows the number of image patches of 501*501 pixels in training and testing set.

Figure 2

See image above for figure legend.

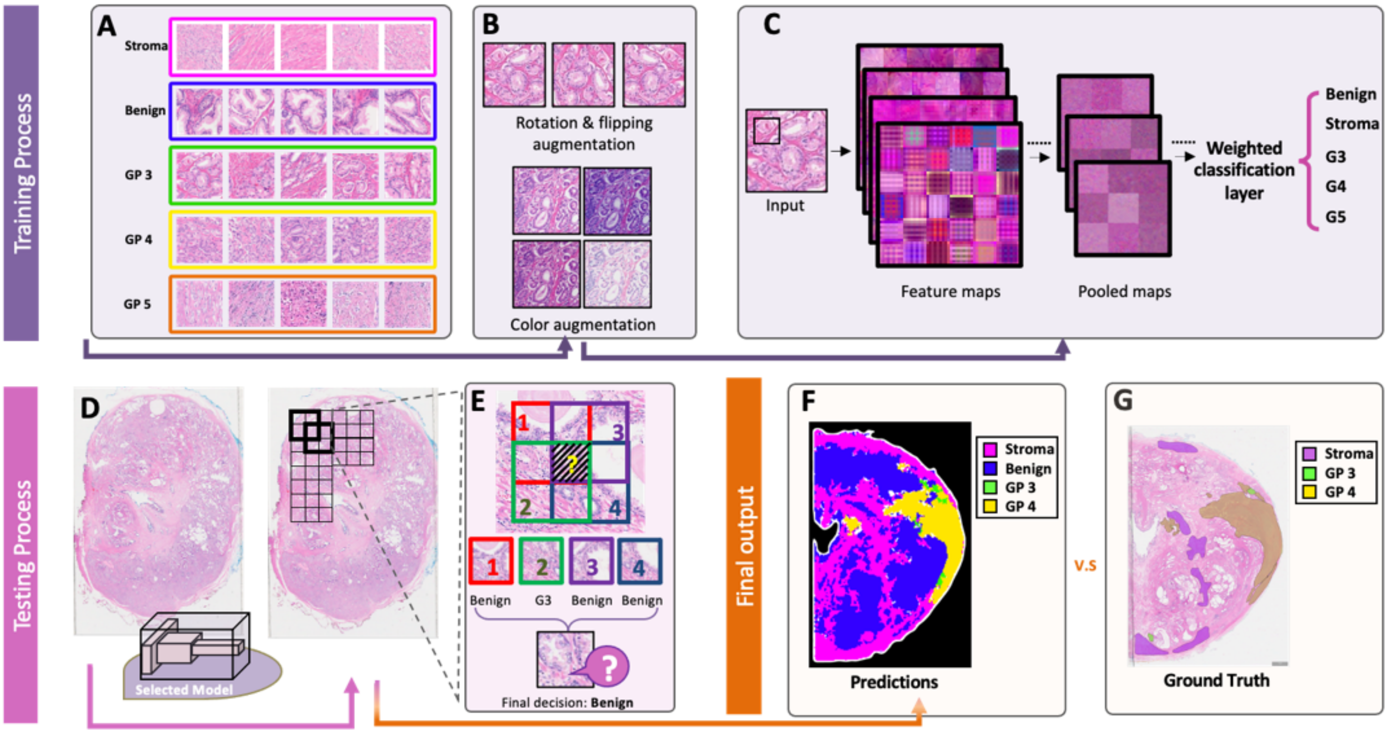

Figure 3 Training and testing of the AI model After organizing the image patches for training, data augmentation was applied to every patch, as demonstrated in **(A)** – **(B)**. Rotation angle is picked randomly from a continuous uniform distribution within 0~360, and each patch is reflected horizontally or vertically with 50% probability. Color augmentation changed the intensity of pixel value randomly within a certain range, and detail is described in supplementary table 4. In this study, structures of ResNet50, VGG16 and NasNet Mobile were used, and their regular classification layers were replaced with the weighted classification layer. Common structures are demonstrated in **(C)**. In the testing process, we applied the trained model to the test image with a sliding window operation with voting strategyas shown in **(D)** and **(E)**. **(F)** – **(G)** show an example of prediction results and its corresponding ground truth.

Figure 3

See image above for figure legend.

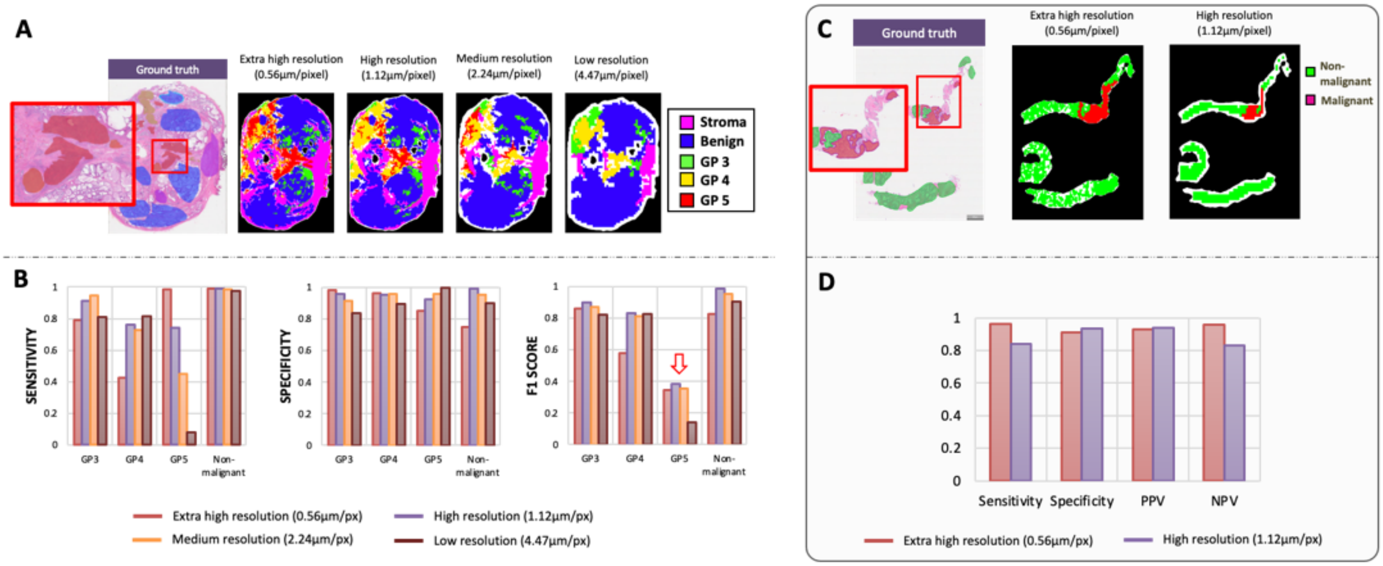

Figure 4 Scale optimization and model performance on prostatectomy and biopsy specimens Model performances at different resolution are significantly different, as shown in **(A).** To quantify the difference, **(B)** shows the comparison of sensitivity, specificity and F1 score. Model of high resolution were therefore selected to process prostatectomy specimens. Similarly, models at different resolutions were applied to biopsy images, as shown in **(C)-(D).** Considering the shape and size of the biopsy, model at extra high resolution is more desirable. In particular, part of the biopsy was barely processed under lower resolution but correctly identified as benign tissue at extra high resolution.

Figure 4

See image above for figure legend.

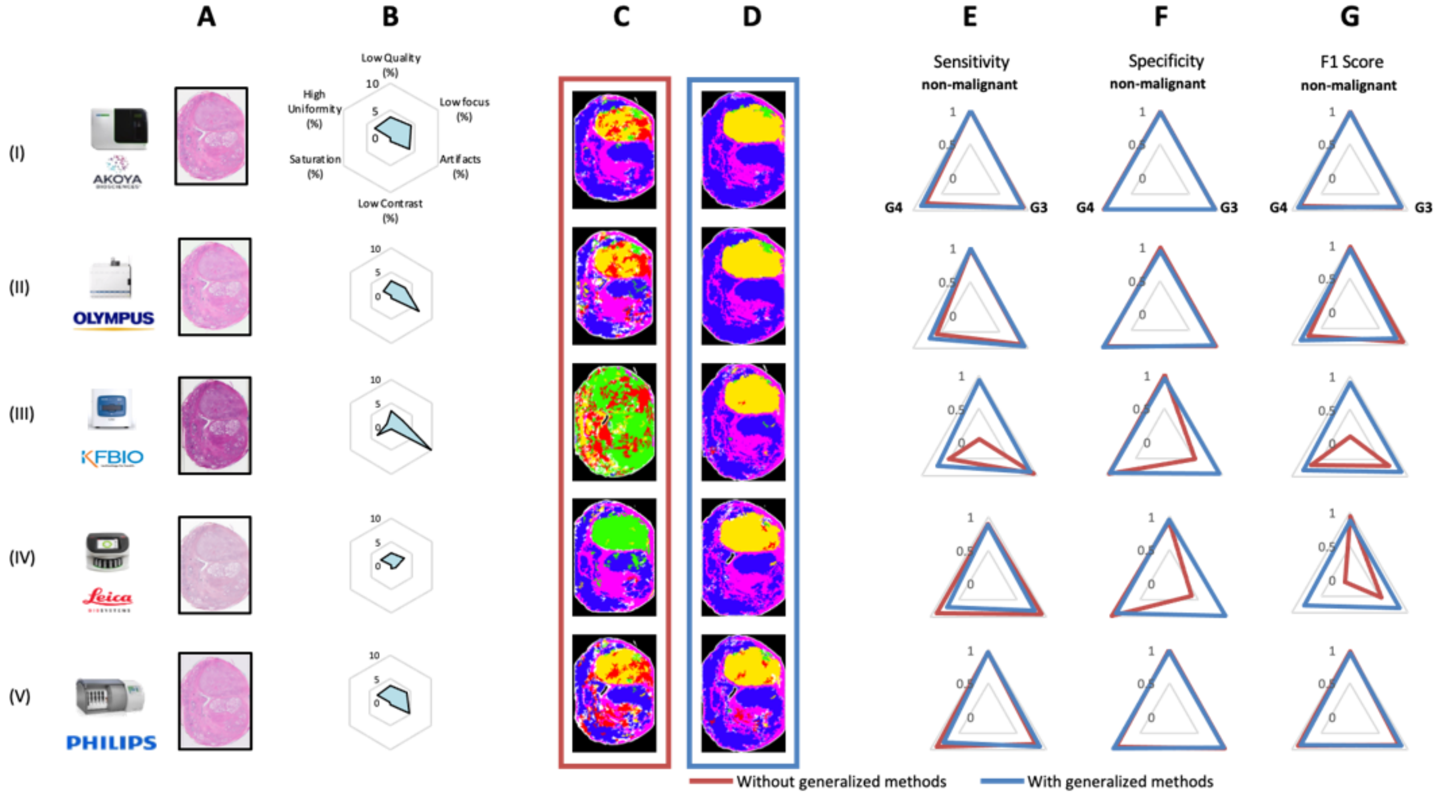

Figure 5 Generalization across images scanned by different scanners (A) shows examples of images acquire from different scanners: (I) Akoya Biosciences, (II) Olympus, (III) KFBio, (IV) Leica, (V). A!magQC assessed the images scanned by different scanners and identify 5 common quality issues. **(B)**The average percentage of image tiles with different quality issues, for different scanner dataset. The "low quality" is a weighted average score of all other 5 features. **(C)** and **(D)** compare the generated heatmaps without and with generalized methods, respectively. **(E), (F), (G)** show the sensitivity, specificity and F1 score of different scanner dataset, without and with generalized method. Gleason Pattern5 is not included because there are almost annotated G5 region in the cases scanned by multiple scanners.

Figure 5

See image above for figure legend.

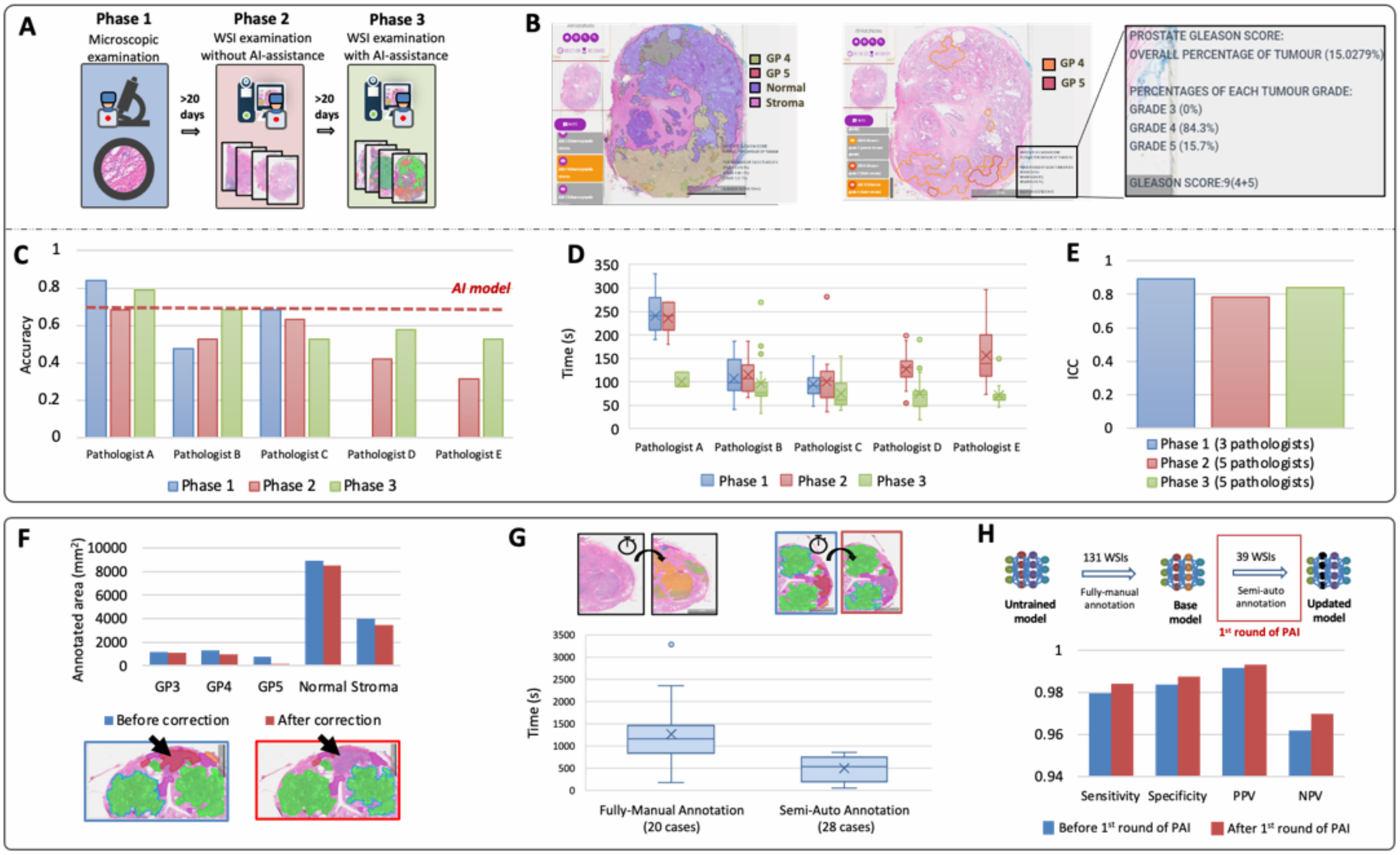

Figure 6 Clinical assessments We designed a 3-phase clinical experiment to assess AI-assisted diagnosis by comparing microscopic examination, WSI examination with and without AI-assistance, as shown in **(A).** The pseudo annotation we presented to pathologists are simplified according to pathologists' feedback, since the original prediction generated by the model is too detailed and contain too much information, as shown in **(B).** In **(C),** we compare the accuracy of Gleason grading between different pathologists at different phase. Our AI-model achieved pathologist-level performance. The examination time is summarized in **(D).** To assess the pathologists' agreement, **(E)** shows the ICC of Gleason grading in different phases. 2 pathologists from China didn't participate in phase 1. As for the PAI experiment, **(F)** Compare the annotation area of each class before and after the pathologist's adjustment, and **(G)** compare the annotation time of fully-manual and semi-automatic methods. Model performance on tumour detection increased after 1st of PAI, as shown in **(H)**

Figure 6

See image above for figure legend.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFinal.pdf