



**HAL**  
open science

## Evaluation of Artificial Intelligence-Based Gleason Grading Algorithms "in the Wild"

Khrystyna Faryna, Leslie Tessier, Juan Retamero, Saikiran Bonthu, Pranab Samanta, Nitin Singhal, Solène-Florence Kammerer-Jacquet, Camelia Radulescu, Vittorio Agosti, Alexandre Collin, et al.

► **To cite this version:**

Khrystyna Faryna, Leslie Tessier, Juan Retamero, Saikiran Bonthu, Pranab Samanta, et al.. Evaluation of Artificial Intelligence-Based Gleason Grading Algorithms "in the Wild". *Modern Pathology*, 2024, 37 (11), pp.100563. 10.1016/j.modpat.2024.100563 . hal-04721197

**HAL Id: hal-04721197**

**<https://hal.science/hal-04721197v1>**

Submitted on 4 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

## Research Article

## Evaluation of Artificial Intelligence-Based Gleason Grading Algorithms “in the Wild”

Khrystyna Faryna<sup>a,\*</sup>, Leslie Tessier<sup>a</sup>, Juan Retamero<sup>b</sup>, Saikiran Bonthu<sup>c</sup>, Pranab Samanta<sup>c</sup>, Nitin Singhal<sup>c</sup>, Solene-Florence Kammerer-Jacquet<sup>d</sup>, Camelia Radulescu<sup>e</sup>, Vittorio Agosti<sup>f</sup>, Alexandre Collin<sup>g</sup>, Xavier Farre<sup>h</sup>, Jacqueline Fontugne<sup>i</sup>, Rainer Grobholz<sup>j</sup>, Agnes Marije Hoogland<sup>k</sup>, Katia Ramos Moreira Leite<sup>l</sup>, Murat Oktay<sup>m</sup>, Antonio Polonia<sup>n</sup>, Paromita Roy<sup>o</sup>, Paulo Guilherme Salles<sup>p</sup>, Theodorus H. van der Kwast<sup>q</sup>, Jolique van Ipenburg<sup>a</sup>, Jeroen van der Laak<sup>a,r</sup>, Geert Litjens<sup>a</sup>

<sup>a</sup> Radboud University Medical Center, Computational Pathology Group, Nijmegen, The Netherlands; <sup>b</sup> Paige, New York, New York; <sup>c</sup> AIRA Matrix, Thane, India; <sup>d</sup> Department of Pathology, Rennes University Hospital, Rennes, France; <sup>e</sup> Department of Pathological Anatomy and Cytology, Hopital Foch, Suresnes, France; <sup>f</sup> Department of Medicine and Surgery, University of Brescia, Brescia, Italy; <sup>g</sup> Department of Cell and Tissue Pathology, Angers University Hospital Center, Angers, France; <sup>h</sup> Public Health Agency of Catalonia, Lleida, Spain; <sup>i</sup> Department of Pathology, Institut Curie, Saint-Cloud, France; <sup>j</sup> Institute of Pathology, Cantonal Hospital Aarau, Aarau, Switzerland; <sup>k</sup> Department of Pathology, Isala Zwolle, Zwolle, The Netherlands; <sup>l</sup> Department of Surgery, Faculty of Medicine of the University of Sao Paulo, Sao Paulo, Brazil; <sup>m</sup> Department of Pathology, Memorial Hospitals Group, Istanbul, Turkey; <sup>n</sup> Department of Pathology, Ipatimup, Porto, Portugal; <sup>o</sup> Department of Pathology, Tata Medical Center, Kolkata, India; <sup>p</sup> Teaching and Research Center, Instituto Mario Penna, Belo Horizonte, Brazil; <sup>q</sup> Department of Anatomic Pathology, University Health Network and Princess Margaret Cancer Center, Toronto, Canada; <sup>r</sup> Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden

## ARTICLE INFO

## Article history:

Received 26 December 2023

Revised 4 June 2024

Accepted 9 July 2024

Available online 16 July 2024

## Keywords:

artificial intelligence  
computational pathology  
deep learning  
Gleason grading

## ABSTRACT

The biopsy Gleason score is an important prognostic marker for prostate cancer patients. It is, however, subject to substantial variability among pathologists. Artificial intelligence (AI)-based algorithms employing deep learning have shown their ability to match pathologists' performance in assigning Gleason scores, with the potential to enhance pathologists' grading accuracy. The performance of Gleason AI algorithms in research is mostly reported on common benchmark data sets or within public challenges. In contrast, many commercial algorithms are evaluated in clinical studies, for which data are not publicly released. As commercial AI vendors typically do not publish performance on public benchmarks, comparison between research and commercial AI is difficult. The aims of this study are to evaluate and compare the performance of top-ranked public and commercial algorithms using real-world data. We curated a diverse data set of whole-slide prostate biopsy images through crowdsourcing containing images with a range of Gleason scores and from diverse sources. Predictions were obtained from 5 top-ranked public algorithms from the Prostate cANcer graDe Assessment (PANDA) challenge and 2 commercial Gleason grading algorithms. Additionally, 10 pathologists (A.C., C.R., J.v.I., K.R.M.L., P.R., P.G.S., R.G., S.F.K.J., T.v.d.K., X.F.) evaluated the data set in a reader study. Overall, the pairwise quadratic weighted kappa among pathologists ranged from 0.777 to 0.916. Both public and commercial algorithms showed high agreement with pathologists, with quadratic kappa ranging from 0.617 to 0.900. Commercial algorithms performed on par or outperformed top public algorithms.

© 2024 THE AUTHORS. Published by Elsevier Inc. on behalf of the United States & Canadian Academy of Pathology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

\* Corresponding author.

E-mail address: [khrystyna.faryna@radboudumc.nl](mailto:khrystyna.faryna@radboudumc.nl) (K. Faryna).

## Introduction

The biopsy Gleason score is the primary tissue-based prognostic marker for prostate cancer patients.<sup>1</sup> However, Gleason grading suffers from significant interobserver and intraobserver variability.<sup>2,3</sup> Artificial intelligence (AI)-based algorithms utilizing deep learning have demonstrated the ability to attain performance levels on par with pathologists when assigning a Gleason score.<sup>4,5</sup> In addition, studies have indicated that pathologists can enhance their Gleason grading performance with the assistance of such AI-based systems.<sup>6,7</sup>

The number of AI-based products for histopathology has rapidly increased over the course of past years. For instance, the number of AI exhibitors at the annual meeting of the United States and Canadian Academy of Pathology and European Congress of Pathology has tripled in the time span of 2018-2023.<sup>8-11</sup>

AI models are often trained using data from a limited number of clinical environments, which can inadvertently introduce biases concerning certain patient groups, image acquisition protocols, and imaging devices.<sup>12</sup> As a result, the lack of generalizability may hinder the real-world effectiveness of AI in histopathology. Moreover, the testing data often originate from clinical environments similar to those where training data were obtained and potentially provide an optimistic estimate of the AI algorithm's final performance in real clinical practice. However, obtaining access to larger diverse data sets to develop and test models is complicated because of regulatory factors.

For an extended period of time, the evaluation and validation of novel research methods were based on the private data sets of the authors' institution, making it unfeasible to conduct a fair and direct comparison of various solutions.<sup>13</sup> The initial endeavors to tackle this issue can be traced back to the late 1990s when the first<sup>14</sup> international comparative evaluation of intermodality brain image registration methods was conducted. To guarantee a fair comparison of the algorithms, participants in the study did not have access to the reference standard until after they had submitted their own results. Later, more biomedical imaging algorithm competitions were organized.<sup>15,16</sup> The first so-called grand challenge<sup>17</sup> in biomedical image analysis was organized in the framework of the 2007 International Conference on Medical Image Computing and Computer-Assisted Intervention. As time passed, research methodologies began to change, resulting in a consistent rise in the annual count of organized challenges.<sup>18</sup>

In recent years, AI-based algorithm challenges, such as Prostate cANcer graDe Assessment (PANDA)<sup>4</sup> and cANcer METastases in LYmph nODEs challeNge (CAMELYON),<sup>19</sup> have played a pivotal role in fostering the development and validation of cutting-edge algorithms for various tasks in pathology. Concurrently, the landscape of regulatory approvals for medical software has evolved, with companies<sup>20</sup> actively pursuing Food and Drug Administration (FDA) and Conformité Européenne (CE) approval for their algorithmic solutions. Notably, these academic and commercial spheres, although aligned in their overarching goal of algorithm development and validation, often operate in distinct parallel universes. Academic algorithms are often evaluated on public benchmarks. Commercial algorithms are evaluated within trials, the data from which have limited accessibility to the academic community. Moreover, commercial entities infrequently participate in academic challenges, citing various reasons ranging from licensing constraints to concerns over potential underperformance. The present study aims to bring together these distinct tracks by explicitly comparing the performance of leading

academic challenge algorithms and top commercial algorithms on a single, diverse, and challenging crowdsourced data set.

First, we have curated a data set of whole-slide images (WSIs) of prostate biopsies obtained through crowdsourcing. We utilized social media and personal contacts to engage pathologists from various centers and collect 138 anonymized prostate biopsy slides. Following a quality review process, we curated a final set of 113 cases from 7 different sources, providing a diverse and comprehensive data set that mimics real-world clinical scenarios. Furthermore, the data set exhibits a high variability in scanning and staining protocols. This diversity challenges the AI algorithms to generalize and adapt effectively to varying conditions, mimicking the complexity encountered in actual clinical practice. Second, we asked pathologists from around the world to grade this set of slides. Third, we obtained predictions from top-performing publicly available AI-based Gleason grading algorithms (the winning solutions of the PANDA challenge<sup>4</sup>) on this data set. Finally, we reached out to commercial providers of AI-based Gleason grading algorithms, 2 of which provided entries of their official product predictions on the collected data. We additionally extended the possibility of commercial AI-based Gleason grading vendors to obtain their scores on these data through the biomedical imaging challenge platform [gleason-grading-in-the-wild.grand-challenge.org](https://gleason-grading-in-the-wild.grand-challenge.org).

## Materials and Methods

### Slide Crowdsourcing

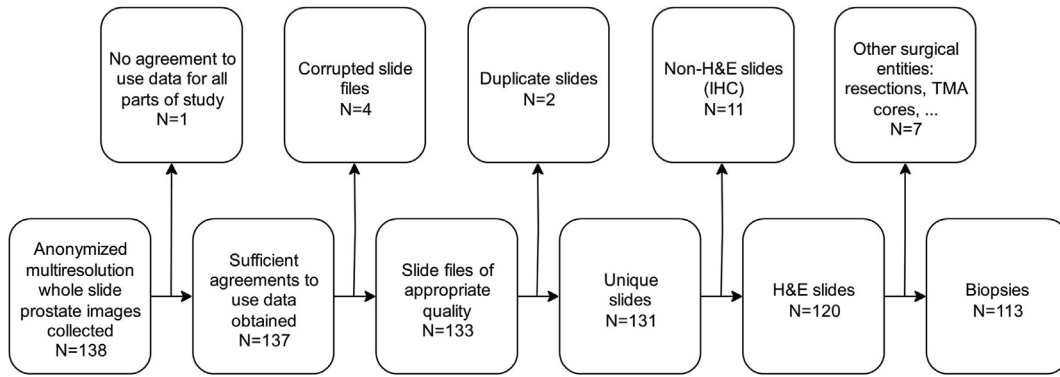
A call for slide collection was posted on social media; additionally, pathologists from various centers around the world were contacted. We requested anonymized prostate biopsy slides from their routine clinical practice.

We initially received 138 slides from 8 sources. The received slides were reviewed by a pathology expert, who checked whether the slides (1) contained prostate tissue, (2) were biopsies, (3) were stained with hematoxylin and eosin, and (4) were of sufficient quality for grading. The slides that did not satisfy the above criteria were excluded. The excluded slides were prostate resections, were immunohistochemically stained slides, or had insufficient quality to perform grading, or the data sharing policy did not allow the use of slides for all parts of the study (see Fig. 1 for details).

The final set consisted of 113 cases. Most of the cases included multiple biopsies. A pathology resident under a subspecialized uropathologist's supervision reviewed the slide's original grade and selected 1 representative biopsy per case.

### Data Diversity

The final data set consisted of 113 cases from 7 sources. The data set included data from 5 scanners: 3DHISTECH P1000, Hamamatsu, Leica Aperio, Philips, and KFBIO. We obtained slides in 5 different data formats: "tiff," "svs," "mrxs," "ndpi," and "kfb." The color profile of the slides was not specified; thus, no color profile correction was performed. The obtained slides had varying original minimal pixel sizes: 0.23, 0.24, 0.25, 0.46, 0.48, and 0.5  $\mu\text{m}$  per pixel. The original slides had either red, green, blue or red, green, blue, alpha channel format. Some slides included artifacts, such as colored ink, markers, and foreign tissue, from the colon. All slides used in this study were stained with hematoxylin and eosin; in addition, the staining of the slides obtained from institutions in



**Figure 1.** Slide inclusion pipeline. H&E, hematoxylin and eosin; IHC, immunohistochemistry; TMA, tissue microarray.

France included saffron in addition to hematoxylin and eosin. **Figure 2** shows the examples of diversity in the data.

pathology resident (L.T.) under the supervision of a subspecialized uropathologist (J.v.I.).

*Preprocessing*

First, a single representative biopsy, selected by a pathology resident, was cropped from each slide because slides could contain multiple biopsies or different levels of the same biopsy. The obtained slides had heterogeneous original micron per pixel resolution, for example, 0.23, 0.46, 0.92, ... or 0.48, 0.96, 1.92, ...  $\mu\text{m}$  per pixel. Thus, from each slide, we extracted the level closest to 0.5- $\mu\text{m}$  per pixel spacing. The slides were subsequently resampled to have 0.50, 1.00, 2.00, 4.00, 8.00, ...  $\mu\text{m}$  per pixel resolution each. All slides were saved as 3-channel red, green, blue multiresolution slides in a standard pyramidal tiff format.

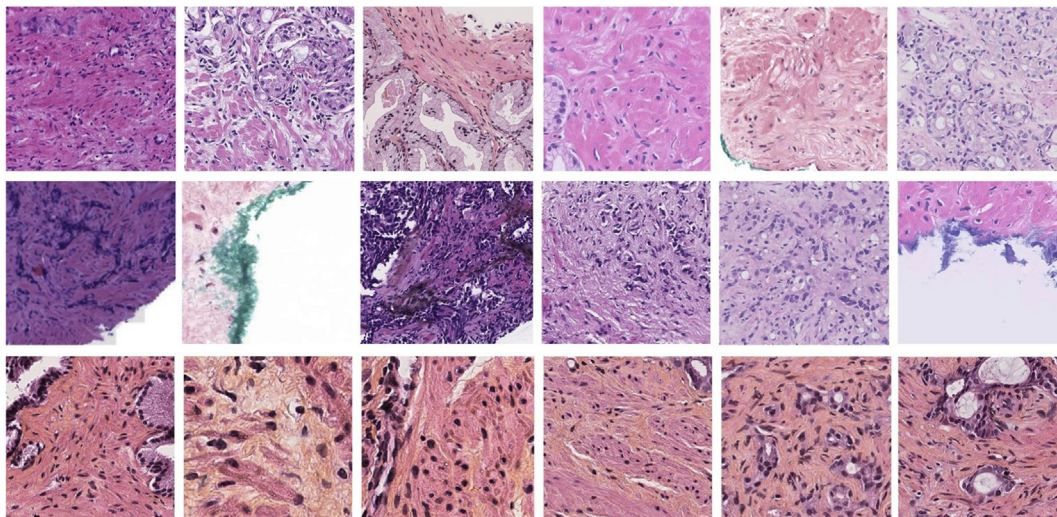
*Reader Study*

The reader study was handled through <https://grand-challenge.org/> platform using a Cirrus Core (<https://www.radboudumc.nl/en/research/radboud-technology-centers/deep-learning/hardware-software/cirrus-core>) viewer. In total, 10 pathologists (A.C., C.R., J.v.I., K.R.M.L., P.R., P.G.S., R.G., S.F.K.J., T.v.d.K., X.F.) graded the slides in a reader study. The participants reported an average of 17.5 and a median of 18.5 years of experience in general pathology. Nine of 10 pathologists who participated in this study were subspecialized in uropathology. The participants reported an average of 14 and a median of 15.5 years of experience in uropathology. The readers were presented with 113 slides containing a single prostate biopsy and were asked to answer 3 questions about each slide:

*Reference Standard*

During the data collection, we requested pathologists from the data-contributing centers to provide a Gleason grade for each case submitted, among other metadata. After selecting a single biopsy per case, the Gleason grade was assigned to each biopsy by a

1. Is there a tumor on this slide, and if so, what Gleason grade would you assign to it? (mandatory)
2. Percentage of majority Gleason pattern? Considering the tumor area, what's the percentage of the most common Gleason



**Figure 2.** Variation among images present in the data set. First row, staining variation; second row, fixation artifacts, ink; and third row, hematoxylin and eosin with saffron staining.



**Figure 3.**

The Gleason grade group score for each case in the data set provided by each algorithm and pathologist in the study. The reference standard is the majority vote of all the pathologists. First, we calculated whether the case was graded as malignant or benign by the majority of pathologists. Second, for malignant cases, a majority grade group was calculated. In the case of ties, the higher value group was assigned.

pattern you assigned to this tumor? For example: If you said Gleason 3+4, with the Gleason 3 pattern representing 70% of the tumor area, enter 70. If you said 3+3 or 4+4, then enter 100. If you said no tumor: enter 0. Please enter only numbers from 0 to 100. (mandatory)

3. Comment. Do you have any comments you want to share on this specific slide? (optional)

### Artificial Intelligence Algorithms

#### Public Artificial Intelligence Algorithms

The PANDA<sup>4</sup> was a Gleason grading AI algorithm challenge that was hosted on <https://kaggle.com> platform in 2019, with >1000 participating teams from around the world. In this study, we use the 5 top-performing algorithms based on generalization performance: NS\_Pathology, PND, BarelyBears, Kiminya, and Vanda. A more detailed description of the algorithms can be found in Bulten et al (2022).<sup>4</sup>

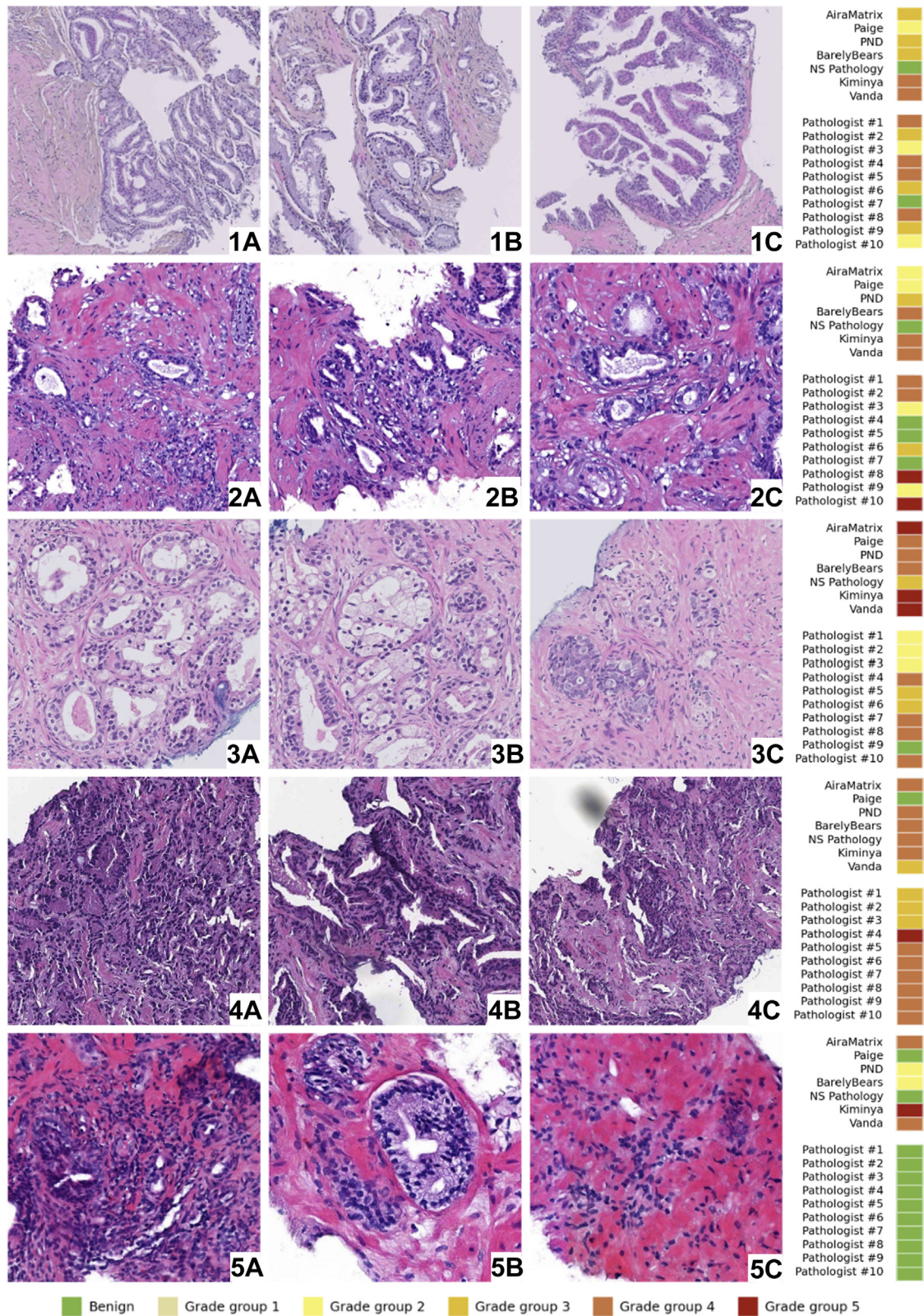
#### Commercial Artificial Intelligence Algorithms

We have published a call for the evaluation of Gleason grading AI-based algorithm on social media. We have also sent personal

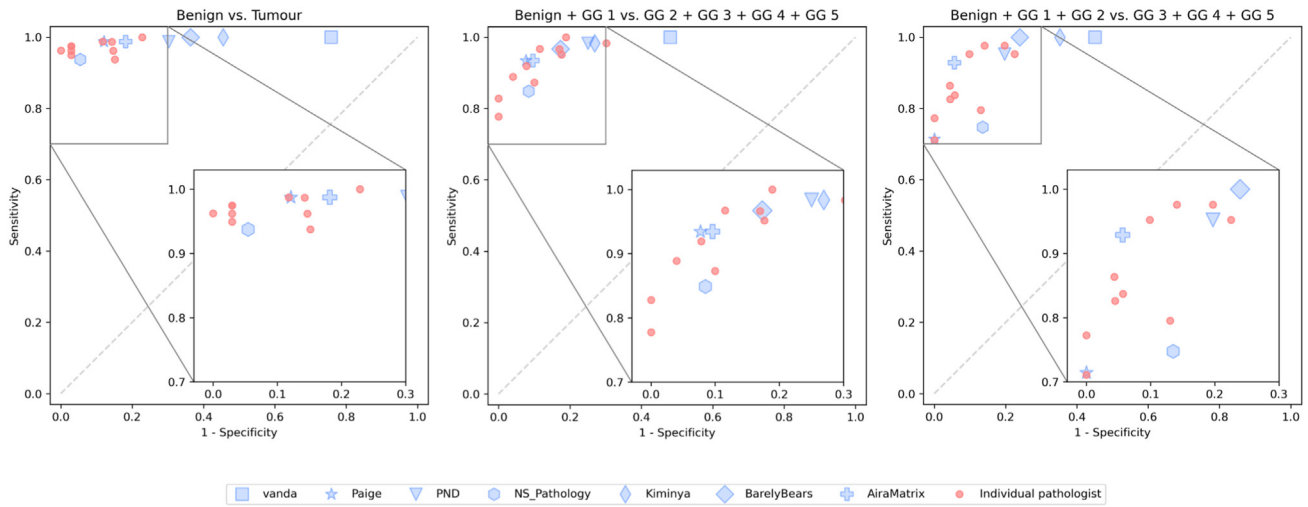
invitations to 3 out of 3 companies we found to have a Conformité Européenne In-Vitro Diagnostic (CE-IVD) marked Gleason grading algorithm at the time of the study launch (beginning of May 2022), of which 1 was accepted. The search for companies was based on publicly available information. At the time of the study launch, there was no central register of CE-marked AI-based products in histopathology. Two companies provided entries of their algorithm predictions for the data: Paige and AIRA Matrix.

#### Paige: Paige Prostate

The details about Paige Prostate (PaPr) AI-based algorithm have been reported elsewhere.<sup>21,22</sup> Briefly, PaPr is a deep learning–based system trained using multiple-instance learning. This weakly supervised approach did not require pixel-level manual annotations. A large data set comprising >32,000 prostate biopsy WSIs from approximately 7000 patients scanned at ×20 magnification was used. The pathology report was the ground truth for each WSI. Slides were prepared and diagnosed by genitourinary pathologists at the Memorial Sloan Kettering Cancer Center, New York, New York. PaPr is made of different modules. PaPr Detection outputs a binary WSI-level classification for suspicion of cancer based on applying a cutoff value to the continuous score, and if a WSI is suspicious, it displays a location



**Figure 4.** Regions of interest from cases with high disagreement among algorithms or pathologists. The grades provided by pathologists and algorithms are shown on the right.



**Figure 5.** Performance of pathologists and algorithms at clinically relevant decision thresholds: left, benign vs tumor; middle, benign + GG 1 vs the rest; and right, benign + GG 1 + GG 2 vs GG 3 + GG 4 + GG 5. The sensitivity and specificity of each single pathologist were calculated against the majority vote of the rest of the pathologists. GG, grade group.

on the WSI with the greatest probability for cancer. This module is, to date, the only AI tool in pathology authorized by the Food and Drug Administration in the United States for clinical use.<sup>22</sup> In addition, PaPr is CE-IVD and United Kingdom Conformity Assessed-approved.

PaPr Grade and Quantify grades and quantifies tumor content on the WSI, providing a slide-level Gleason score. It highlights areas suspicious for cancer by Gleason pattern and determines overall tumor percentage and length. It is CE-IVD and United Kingdom Conformity Assessed-approved. In this study, versions 3.3.7 (Detection) and 3.0.4 (Grade and Quantify) were used.

**AIRA Matrix: AIRAProstate**

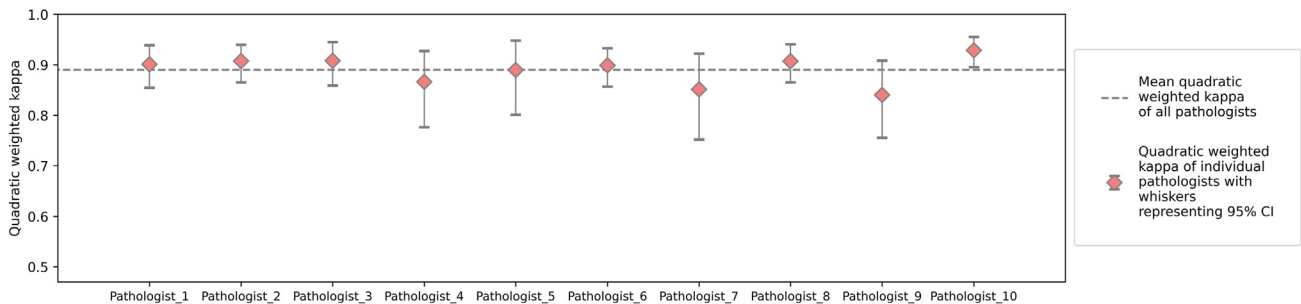
A description of the AIRA Matrix Prostate tumor identification and Gleason grading algorithm (AIRAProstate) has been previously reported.<sup>23</sup> AIRAProstate is an approach for segmenting and grading epithelial tissue that is based on deep learning. The system selects samples for annotation using active learning and uncertainty measures in a semisupervised manner. For increased generalizability, a novel convolutional neural network architecture has been implemented, which learned domain-agnostic features. During training, the system was exposed to >10,000 WSIs from needle core biopsies. Furthermore, it underwent validation on >11,000 cases originating from the United States, Europe, and India. The system offers a range of clinically relevant measurements, including core length, tumor length, percentage of

tumor area, percentage of the area classified as grades 3, 4, and 5 and the International Society of Urologic Pathologists grade group. The system was designed specifically to identify and classify individual tumor glands according to their correspondence with Gleason patterns 3, 4, and 5. In addition, the identification of gland levels facilitates more precise quantification of the tumor and various Gleason patterns, leading to enhanced classification accuracy when distinguishing between the International Society of Urologic Pathologists grade groups 2 and 3.

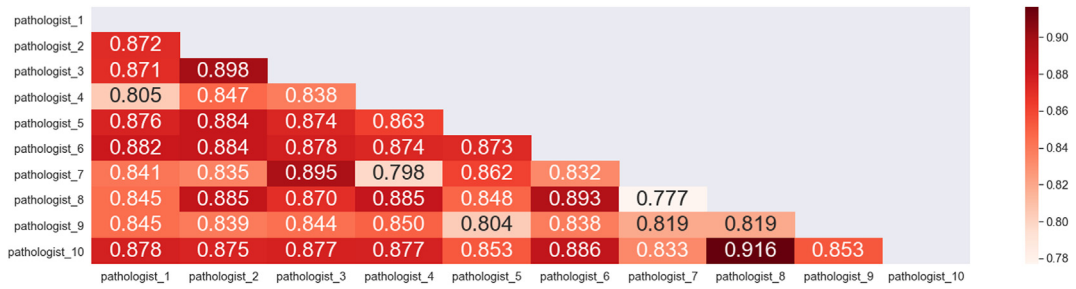
**Evaluation Metrics**

To evaluate the pathologists' performance, we computed the quadratic weighted kappa (QWK) of each pathologist against the majority vote of the rest and the pairwise agreement of pathologists against each other. In addition, the sensitivity and specificity of each pathologist against the majority vote of the rest were computed for clinically relevant thresholds: benign vs tumor, benign + grade group 1 vs the rest, and benign + grade group 1 + grade group 2 vs the rest.

To evaluate AI algorithms, we computed their QWK against the majority vote of all the pathologists. In case of ties, the majority vote output was assigned to the higher Gleason grade group. In addition, the sensitivity and specificity of each algorithm against the majority vote of all the pathologist rest were computed for clinically relevant thresholds: benign vs tumor, benign + grade group 1 vs the rest, and benign + grade group 1 + grade group 2 vs the rest. The



**Figure 6.** Individual agreement (quadratic weighted kappa) of each pathologist with the majority vote of the rest.



**Figure 7.** Pairwise agreement (quadratic weighted kappa) of pathologists against each other.

statistical significance was assessed using a 2-sided permutation test, with .05 as the significance level.

**Results**

The grade for each case, assessed both by pathologists and by public and commercial algorithms, is shown in Figure 3. Overall, there is a high agreement between algorithms and pathologists.

A pathology resident assessed the slides that had a high disagreement among pathologists or algorithms and selected regions that could have been a possible source of confusion. Figure 4 shows regions from the slides with high disagreement among algorithms or pathologists. For case 1 in Figure 4, there was a high disagreement both among pathologists and algorithms, the possible sources of variation could be the resemblance of the cribriform pattern (1A) and high-grade prostatic intraepithelial neoplasia (1C). This slide (1A, 1B, and 1C) is also stained with saffron in addition to hematoxylin and eosin. Case 2 in Figure 4 has a crowdedness of small glands (2C) and growth patterns that can be misinterpreted for a higher Gleason pattern (2A or 2B). The original clinical conclusion for the whole case was reported as grade group 2. Case 3 was graded higher by algorithms than by pathologists. This case is an uncommon variant of prostate cancer called “foamy gland adenocarcinoma”; this variant has a high density of glands with minimal cytologic atypia (3B) and a basal layer difficult to evaluate (3A). Some areas could have potentially been misinterpreted as Gleason 5 (3C) by an algorithm. Case 4 has considerable fixation and stretch artifacts (4C), tissue folds (4B), and, likely, a higher tissue thickness that resulted in a darker stained slide. Case 5 was unanimously graded as benign by all pathologists but caused a high disagreement among algorithms. This case includes inflammation and hypercellular areas (5A, 5B,

and 5C). Pathologists reach higher specificity when distinguishing between tumor and benign cases (Fig. 5, left); algorithms, in turn, reach a higher sensitivity score (Fig. 5, left).

*Reader Study*

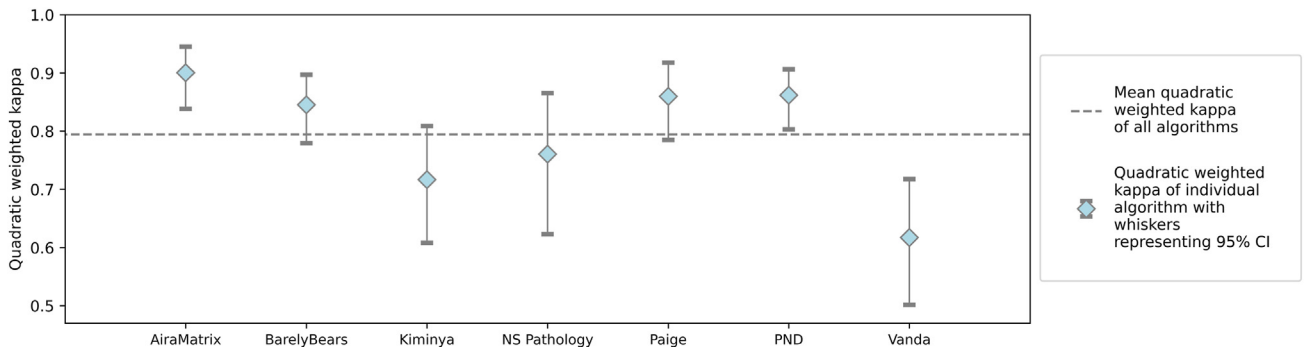
The QWK of each pathologist against the majority vote of the rest is shown in Figure 6, and the pairwise agreement of pathologists against each other is shown in Figure 7. The average QWK of pathologists against the majority vote of the rest is 0.890, and it ranges from 0.840 to 0.929. The average pairwise agreement of pathologists is 0.858, and it ranges from 0.777 to 0.916.

*Artificial Intelligence Algorithms*

To evaluate the performance of AI algorithms, we first compute a majority vote of pathologists for each case. In case of ties, the higher Gleason grade is selected. We compute the agreement (QWK) of algorithms with the majority vote of pathologists (Fig. 8) and the performance of the algorithms at clinically relevant thresholds (Fig. 5): identifying benign cases vs tumor, benign + grade group 1 vs the rest, and benign + grade group 1 + grade group 2 vs the rest.

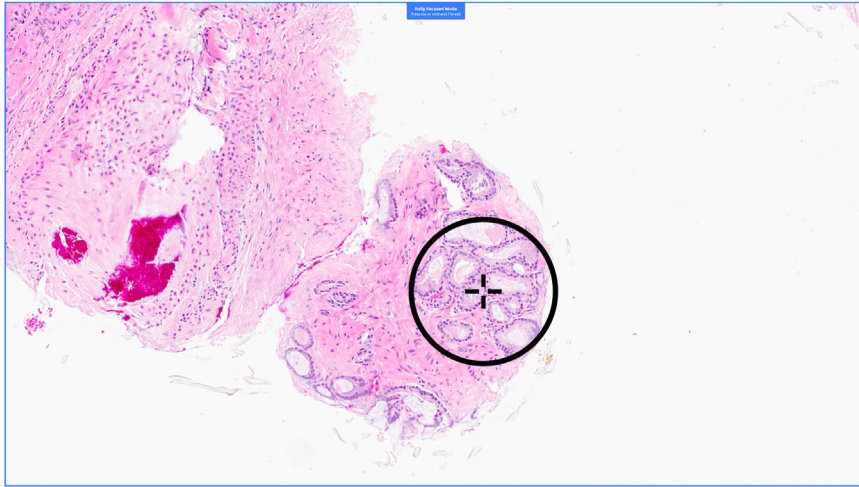
*Public Artificial Intelligence Algorithms*

Among public AI algorithms, PND, BarelyBears, and NS\_Pathology show better performance in comparison to the rest, with their QWK against pathologists being 0.862, 0.845, and 0.760, respectively (Fig. 8). NS\_Pathology archives the highest specificity



**Figure 8.** Individual agreement (quadratic weighted kappa) of algorithms with the majority vote of pathologists.



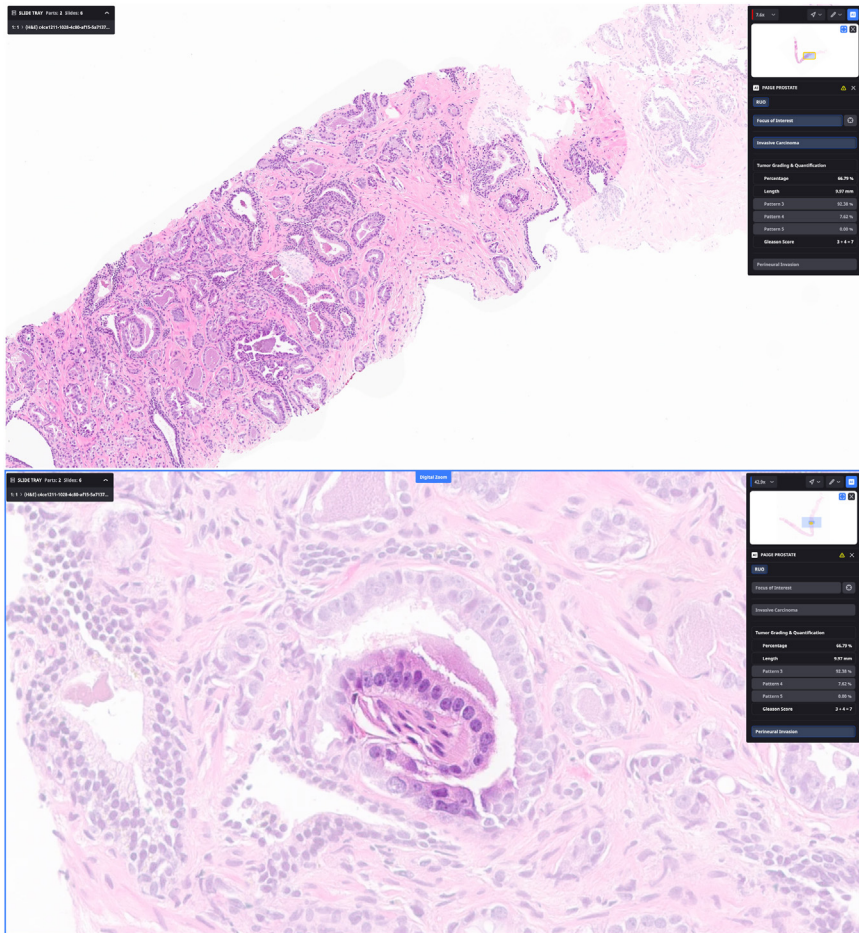


**Figure 9.**

The Food and Drug Administration–approved Paige Prostate Detection is a binary detector that classifies WSIs as either benign or suspicious for cancer. When cancer is detected on a WSI, a crosshair indicates the location with the greatest probability for cancer. This is helpful when assessing WSIs with low tumoral burden. WSI, whole-slide image.

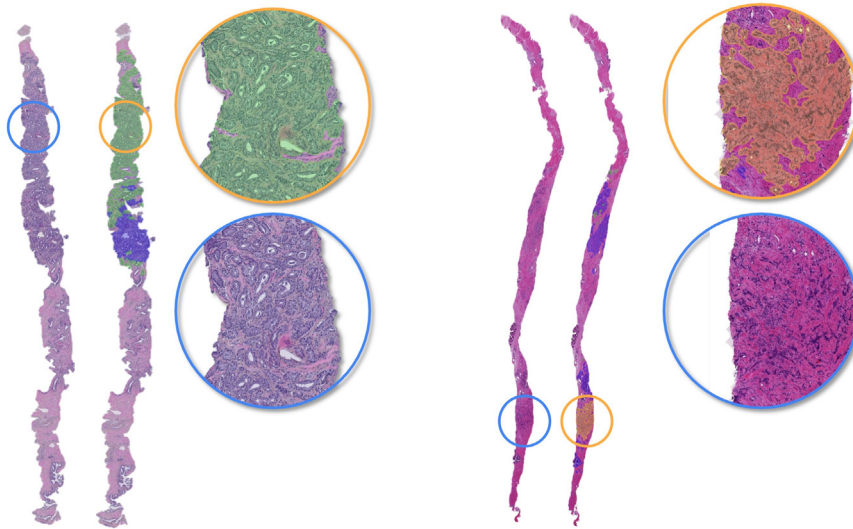
of 0.945 among all algorithms at a sensitivity of 0.938 in tumor vs benign detecting threshold, outperforming both public and commercial algorithms in distinguishing tumor vs benign prostate

tissue (Fig. 5, left). At the clinically relevant threshold of benign + grade group 1 vs the rest, NS\_Pathology, PND, and BarelyBears reach specificities of 0.915, 0.750, and 0.827 at sensitivities of



**Figure 10.**

Paige Grade and Quantify displays all regions suspicious for cancer by reducing the contrast of the benign areas. The pathologist's attention is thus directed to the regions of interest (top). In addition, a side panel (right-hand side) displays the overall tumor percentage present on the slide, tumor length, and a breakdown of the different Gleason patterns present and their corresponding percentages. An overall Gleason score is also given. Perineural invasion is also detected and displayed (bottom).



**Figure 11.** The AIRAPProstate system identifies and classifies tumor glands subsequently assigning them to one of the Gleason patterns.

0.849, 0.984, and 0.967, respectively (Fig. 5, middle). At the clinically relevant threshold of benign + grade group 1 + grade group 2 vs the grade group 3 + grade group 4 + grade group 5, PND and

BarelyBears reach specificities of 0.803 and 0.761 at sensitivities of 0.952 and 1.000 outperforming other public algorithms (Fig. 5, right).



**Figure 12.** The AIRAPProstate software generates vector annotations that are visually represented in different colors, each relating to a certain clinically significant parameter.

## Commercial Artificial Intelligence Algorithms

Overall, commercial AI algorithms perform either on par or outperform the public ones. On this data set, the QWK of AIRA Matrix and Paige against the majority vote of pathologists is 0.900 and 0.860, respectively (Fig. 8). The difference between QWK values of PaPr and AIRAProstate is not statistically significant ( $P = .326$ ). At the tumor vs benign detection threshold, Paige reaches a specificity of 0.879 at a sensitivity of 0.988, whereas AIRA Matrix reaches a specificity of 0.918 at a sensitivity of 0.988 (Fig. 5, left). At the clinically relevant threshold of benign + grade group 1 vs the rest, Paige reaches a specificity of 0.923 at a sensitivity of 0.934, whereas AIRA Matrix reaches a specificity of 0.904 at a sensitivity of 0.934 (Fig. 5, middle). At the clinically relevant threshold of benign + grade group 1 + grade group 2 vs the grade group 3 + grade group 4 + grade group 5, AIRA Matrix reaches a specificity of 0.944 at a sensitivity of 0.929, whereas Paige reaches a specificity of 1.000 at a sensitivity of 0.714 (Fig. 5, right). The visual results of PaPr and AIRAProstate prediction are shown in Figures 9 and 10 and Figures 11 and 12, respectively.

## Discussion

To the best of our knowledge, this is the first study providing an estimate of top-performing public and commercial AI-based Gleason grading algorithms on an independent benchmark data set. The data collected via crowdsourcing carry a high degree of variability and thus closely resemble real-world data. This challenge aims to assess the generalizability of AI-based Gleason grading algorithms by testing their performance on data sets that may significantly differ from those used during their initial development. The ability to generalize is crucial for ensuring optimal performance across diverse populations, without additional calibration or retraining. Overall, on this data set, both public and commercial top-performing AI-based Gleason grading algorithms have a high agreement with pathologists. The pairwise QWK among pathologists ranges from 0.777 to 0.916. Both public and commercial algorithms have shown a high agreement with pathologists, with the QWK ranging from 0.600 to 0.908. The highest performance was achieved by commercial algorithms, whereas the top-performing public algorithms achieved comparable performance to the commercial ones.

On average, commercial algorithms have a higher agreement with pathologists in comparison to public ones on these data. Commercial algorithms have a tendency to undergrade cases in comparison to academic ones, the possible reason behind this is that algorithms from the PANDA challenge were specifically optimized for kappa, whereas commercial algorithms were optimized for clinical decision-making.

It is important to mention that the intended use of AI-based Gleason grading algorithms presented in this study is as an adjunct and diagnostic aid to the pathologists and not as a stand-alone diagnostic tool. Therefore, pathologist supervision remains an essential part of the diagnostic equation. To the best of our knowledge, to this date, no stand-alone AI-based algorithms have been approved by relevant authorities for decision-making within histopathology clinical practice.

As for limitations, this research only covers a subset of currently available commercial Gleason grading algorithms. In the scope of this study, we did not investigate AI's performance against pathologists in the context of clinical outcomes for patients or treatment selection through AI-driven approaches. Additionally, it is noteworthy that leading academic algorithms

from the PANDA challenge provide only a single value as an output prediction and, therefore, lack interpretability.

In future research, it would be beneficial to cover a larger number of commercial algorithms. Additionally, there is a growing need for a comprehensive evaluation of the potential impact of AI-based algorithmic grading on improving patient outcomes. To facilitate robust evaluations and comparisons of AI-based algorithms in histopathology, it is imperative to develop extensive, diverse, and multiinstitutional benchmark data sets that can serve as a testing ground for both public and commercial software applications.

The lack of standardization in digital pathology leads to significant variations in the properties of resulting images. These variations encompass aspects, such as image resolution, color calibration, and file formats, making it challenging to ensure consistency and reliability in digital pathology workflows. Standardization efforts are crucial to mitigate these discrepancies, ensuring that pathologists and AI algorithms can consistently and accurately interpret and analyze digital pathology images, ultimately improving patient diagnosis and care.

Although public model performance is often reported on both private and public data, commercial algorithms undergo a certification process primarily based on private data evaluation. Unlike commercial algorithms, public algorithms are commonly assessed through public benchmarks and challenges. There exists a notable absence of comprehensive evaluation that encompasses both public and commercial algorithms. This research paper aims to bridge this gap by providing an independent evaluation of public and commercial algorithms on a common benchmark.

## Acknowledgments

The authors would like to thank the [grand-challenge.org](https://grand-challenge.org) platform for hosting the reader study and the live leaderboard <https://gleason-grading-in-the-wild.grand-challenge.org/>.

## Author Contributions

K.F., G.L., J.v.d.L., and L.T. performed study concept and design. K.F., J.R., S.B., P.S., and N.S. worked on algorithm predictions. L.T. worked on reader study setup and evaluation. S.-F.K.-J., C.R., V.A., A.C., X.F., J.F., R.G., A.M.H., K.R.M.L., M.O., A.P., P.R., P.G.S., T.H.v.d.K., and J.v.I. contributed to data collection or slide evaluation and manuscript writing. All authors contributed to review and revision of the manuscript.

## Data Availability

The data of this project are available only within the study frameworks. To evaluate artificial intelligence-based Gleason grading algorithms on the data, proceed to <https://gleason-grading-in-the-wild.grand-challenge.org/>.

## Funding

The collaboration project is cofunded by the Public-Private Partnerships Allowance made available by Health-Holland (Top-sector Life Sciences & Health), to stimulate public-private partnerships.

## Declaration of Competing Interest

Geert Litjens reports financial support was provided by Dutch Research Council. Geert Litjens reports financial support was

provided by HealthHolland. Geert Litjens reports a relationship with Canon Health Informatics that includes: consulting or advisory. Geert Litjens reports a relationship with Aiosyn that includes: equity or stocks. Juan A. Retamero reports a relationship with Sakura Finetek Europe BV that includes: consulting or advisory. Juan A. Retamero is an employee and stock holder of Paige.ai. Nitin Singhal is an employee of AIRA Matrix. Saikiran Bonthu is an employee of AIRA Matrix. Pranab Samanta is an employee of AIRA Matrix. Jeroen van der Laak reports a relationship with Philips NV, Netherlands that includes: consulting or advisory and funding grants. Jeroen van der Laak reports a relationship with ContextVision, Sweden that includes: consulting or advisory and funding grants. Jeroen van der Laak reports a relationship with Aiosyn, Netherlands that includes: employment and equity or stocks. Jeroen van der Laak reports a relationship with Sectra, Sweden that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Ethics Approval and Consent to Participate

Not applicable.

#### References

- Epstein JI. An update of the Gleason grading system. *J Urol*. 2010;183(2):433–440. <https://doi.org/10.1016/j.juro.2009.10.046>
- Egevad L, Ahmad AS, Algaba F, et al. Standardization of Gleason grading among 337 European pathologists. *Histopathology*. 2013;62(2):247–256. <https://doi.org/10.1111/his.12008>
- Allsbrook WC Jr, Mangold KA, Johnson MH, Lane RB, Lane CG, Epstein JI. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Hum Pathol*. 2001;32(1):81e88. Published correction appears in *Hum Pathol*. 2001;32(12):1417. <https://doi.org/10.1053/hupa.2001.21135>
- Bulten W, Kartasalo K, Chen PC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med*. 2022;28(1):154–163. <https://doi.org/10.1038/s41591-021-01620-2>
- Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21(2):233–241. [https://doi.org/10.1016/S1470-2045\(19\)30739-9](https://doi.org/10.1016/S1470-2045(19)30739-9)
- Bulten W, Balkenhol M, Belinga JA, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol*. 2021;34(3):660–671. <https://doi.org/10.1038/s41379-020-0640-y>
- Steiner DF, Nagpal K, Sayres R, et al. Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies. *JAMA Netw Open*. 2020;3(11):e2023267. <https://doi.org/10.1001/jamanetworkopen.2020.23267>
- Annual meeting of United States and Canadian Academy of Pathology (USCAP). Accessed November 23, 2023. [https://www.uscap.org/public/documents/2019-Annual-Meeting/USCAPAM2019\\_Sponsor\\_Exhibitor\\_Information.pdf](https://www.uscap.org/public/documents/2019-Annual-Meeting/USCAPAM2019_Sponsor_Exhibitor_Information.pdf)
- Annual meeting of United States and Canadian Academy of Pathology (USCAP). Accessed November 23, 2023. <https://s36.a2zinc.net/clients/aimusa/uscap2023/Public/Exhibitors.aspx>
- 30th European Congress of Pathology (ECP). Accessed November 23, 2023. [http://www.cpo-media.net/ECP/2018/ECP2018\\_FinalProgramme/HTML/182/](http://www.cpo-media.net/ECP/2018/ECP2018_FinalProgramme/HTML/182/)
- 36th European Congress of Pathology. Accessed November 23, 2023. <https://www.esp-congress.org/sponsors/acknowledgements2023.html>
- Ahluwalia M, Abdalla M, Sanayei J, et al. The subgroup imperative: chest radiograph classifier generalization gaps in patient, setting, and pathology subgroups. *Radiol Artif Intell*. 2023;5(5):e220270. <https://doi.org/10.1148/ryai.220270>
- Price K. Anything you can do, I can do better (no you can't).... *Comput Vis Graph Image Process*. 1986;36:387–391. [https://doi.org/10.1016/0734-189X\(86\)90083-6](https://doi.org/10.1016/0734-189X(86)90083-6)
- West J, Fitzpatrick JM, Wang MY, et al. Comparison and evaluation of retrospective intermodality brain image registration techniques. *J Comput Assist Tomogr*. 1997;21(4):554–566. <https://doi.org/10.1097/00004728-199707000-00007>
- Kalpathy-Cramer J, De Herrera AGS, Demner-Fushman D, Antani S, Bedrick S, Müller H. Evaluating performance of biomedical image retrieval systems—an overview of the medical image retrieval task at ImageCLEF 2004–2013. *Comput Med Imaging Graph*. 2015;39:55–61. <https://doi.org/10.1016/j.compmedimag.2014.03.004>
- Müller H, Rosset A, Vallée J-P, Terrier F, Geissbühler A. A reference data set for the evaluation of medical image retrieval systems. *Comput Med Imaging Graph*. 2004;28(6):295–305. <https://doi.org/10.1016/j.compmedimag.2004.04.005>
- Heimann T, van Ginneken B, Styner MA, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans Med Imaging*. 2009;28(8):1251–1265. <https://doi.org/10.1109/TMI.2009.2013851>
- Maier-Hein L, Eisenmann M, Reinke A, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun*. 2018;9(1):5217. <https://doi.org/10.1038/s41467-018-07619-7>. Published correction appears in *Nat Commun*. 2019;10(1):588. <https://doi.org/10.1038/s41467-019-08563-w>
- Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience*. 2018;7(6):giy065. <https://doi.org/10.1093/gigascience/giy065>
- da Silva LM, Pereira EM, Salles PG, et al. Independent real-world application of a clinical-grade automated prostate cancer detection system. *J Pathol*. 2021;254(2):147–158. <https://doi.org/10.1002/path.5662>
- Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309. <https://doi.org/10.1038/s41591-019-0508-1>
- Raciti P, Sue J, Retamero JA, et al. Clinical validation of artificial intelligence-augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. *Arch Pathol Lab Med*. 2023;147(10):1178–1185. <https://doi.org/10.5858/arpa.2022-0066-OA>
- Singhal N, Soni S, Bonthu S, et al. A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Sci Rep*. 2022;12(1):3383. <https://doi.org/10.1038/s41598-022-07217-0>