Artificial intelligence as a
digital fellow in pathology

Human-machine synergy for
improved prostate cancer diagnosis

Wouter Bulten
Artificial intelligence as a digital fellow in pathology: Human-machine synergy for improved prostate cancer diagnosis

The research described in this thesis was carried out at the Computational Pathology Group and the Diagnostic Image Analysis Group, Radboud Institute for Health Sciences, Radboud University Medical Center (Nijmegen, the Netherlands). This work was supported by a grant from the Dutch Cancer Society (KWF), grant number KUN 2015-7970. Financial support for publication of this thesis was kindly provided by Radboud University, Nijmegen.

Typesetting: This book was typeset by the author using \LaTeX
Cover design by: Studio Vandaar, Ede
Printed by: Ipskamp Printing, Enschede
ISBN: 978-94-6421-543-4

© Wouter Bulten, 2021
All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.
Artificial intelligence as a
digital fellow in pathology

Human-machine synergy for
improved prostate cancer diagnosis

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op vrijdag 28 januari 2022
om 10:30 uur precies

door

Wouter Bulten

geboren op 9 augustus 1991
te Nijmegen
Promotoren: Dr. J.A.W.M. van der Laak
Prof. dr. B. van Ginneken

Copromotoren: Dr. G.J.S. Litjens
Dr. C.A. Hulsbergen-van de Kaa
(Laboratorium voor Pathologie Oost-Nederland)

Manuscriptcommissie: Prof. dr. J.A. Schalken
Prof. dr. J. Pluim
(Technische Universiteit Eindhoven)
Dr. J.P.M. Sedelaar
“A computer is an educational device, it is in fact a direct reflection of your own imagination, your own intelligence. And once you’re given the freedom in which to create things and see the immediate response on the screen, then it becomes a very enjoyable experience; you go on to involve yourself in many other things.”

From The Information Society, documentary by David Hoffman, 1979; incorporated in the song Youth by The Midnight, 2018.
# TABLE OF CONTENTS

1 **General introduction** | 1
1.1 Overview & epidemiology | 2
1.2 Prostate cancer | 3
1.3 Gleason grading system | 6
1.4 Digital & computational pathology | 9
1.5 Computer-aided diagnosis | 10
1.6 Deep learning | 12
1.7 Challenges for computational pathology | 14
1.8 Aims and objectives | 18
1.9 Thesis outline | 18

2 **IHC-based epithelium segmentation** | 21
2.1 Introduction | 23
2.2 Related work | 24
2.3 Materials | 26
2.4 Methods | 30
2.5 Results | 33
2.6 Discussion | 35
2.7 Data availability | 39
2.8 Acknowledgements | 39

3 **Automated Gleason grading of prostate biopsies** | 41
3.1 Introduction | 43
3.2 Methods | 44
3.3 Results | 50
3.4 Discussion | 57
3.5 Acknowledgements | 59
3.6 Supplementary figures | 61
3.7 Supplementary methods | 64

4 **The PANDA Challenge** | 69
4.1 Introduction | 71
4.2 Results | 72
4.3 Discussion | 80
4.4 Methods | 87
4.5 Acknowledgements | 90
5 AI-assisted Gleason grading
5.1 Introduction ............................................. 119
5.2 Materials and methods .................................. 120
5.3 Results .................................................. 126
5.4 Discussion .............................................. 133
5.5 Acknowledgements ...................................... 136

6 General discussion ........................................ 137
6.1 Overview ................................................ 138
6.2 Fine-grained segmentations .............................. 138
6.3 AI-powered Gleason grading ............................ 140
6.4 The PANDA Challenge ................................ 141
6.5 The added value of AI-assistance ....................... 142
6.6 On the importance of data .............................. 143
6.7 Between now and clinical practice ..................... 144
6.8 Beyond human reference standards ................... 146
6.9 Conclusion .............................................. 149

Summary .................................................. 151

Samenvatting .............................................. 155

Publications .................................................. 159

Research Data Management .............................. 165

PhD Portfolio ............................................... 167

Bibliography ............................................... 169

Acknowledgements ....................................... 183

Curriculum Vitae ......................................... 187
General introduction
1.1 Overview & epidemiology

On average, one in nine men gets diagnosed with prostate cancer during their lifetime. With more than 1.2 million new cases each year, prostate cancer is the second most common cancer among men. The incidence of prostate cancer is steadily increasing, and as prostate cancer is mostly associated with advancing age, the WHO expects that there will be 2.3 million new patients yearly by 2040.

In the Netherlands specifically, the statistics show the same trend. Yearly, around 14,000 men are diagnosed with prostate cancer for the first time, making it the most common cancer among men in the Netherlands. Compared to only 4,000 new cases in 1989, prostate cancer incidence is increasing steadily due to an aging population and an increase in testing (Figure 1.1).

With an estimate of 350,000 deaths per year globally, prostate cancer has a high incidence-to-mortality ratio and co-occurs with an increased risk of overdiagnosis and overtreatment. Over time, the mortality rate decreased, but we still see a rise in absolute mortality due to a sharp increase in prostate cancer incidence.

Figure 1.1: Absolute incidence (left) and mortality (right) of prostate cancer within the Netherlands from 1989 to 2017.

The risks of overdiagnosis and overtreatment stress the need for an accurate diagnosis of prostate cancer. The key to diagnosing prostate cancer is the histological assessment of prostate tissue by pathologists. Unfortunately, this process suffers from considerable inter- and intraobserver variability. Artificial intelligence (AI) systems could assist pathologists in diagnosing prostate cancer due to their ability to learn patterns from data effectively. This thesis will cover the development, validation, and evaluation of the potential of AI algorithms to improve prostate cancer diagnosis in histological tissue sections (Figure 1.2).
1.2 Prostate cancer

1.2.1 Prostate anatomy

The prostate is a small organ located in the male pelvis (Figure 1.3, left), just below the bladder, and surrounds the urethra. The prostate is part of the male reproductive system, and its function is to secrete a fluid that will contribute to the total volume of the semen and prolong the spermatozoa’s lifespan. Around 30% of the semen consists of prostatic fluid. Muscle cells within the prostate help with the secretion of the semen during ejaculation.

In pathology, the prostate is most commonly divided into three distinct zones: the peripheral, central, and transition zone (Figure 1.3, right). The peripheral zone is the largest, with around 70% of the glandular tissue, and resides at the back of the prostate. The large majority of prostate cancers originate from the peripheral zone.\(^6\,7\)

The central zone consists of around 25% of the tissue and surrounds the urethra. Despite its size, the central zone only accounts for around 2.5% of prostate cancers.\(^6\)

Finally, the transition zone occupies approximately 20% of the tissue and 20% of the found prostate cancers.\(^6\,8\)

1.2.2 Current diagnostic pathway

The possible presence of prostate cancer can be detected using a blood test, which measures the concentration of prostate-specific antigen (PSA) in the blood. PSA is mainly secreted by the epithelial cells within the prostate, and an increase in the number of cells will result in an elevated PSA level. Unfortunately, a high PSA value is not specific to prostate cancer as benign disorders can also elevate the PSA concentration.
Due to the low specificity, screening for prostate cancer using PSA measurements is highly controversial. As an illustration, results of the European Randomized Study of Screening for Prostate Cancer showed that, after a median follow-up of nine years, to save one life, 1410 men would need to be screened, and 48 additional cases of prostate cancer would need to be treated. These numbers become more favorable with longer follow-up times.\textsuperscript{5,10} 

So, while PSA screening could result in a significant decrease in mortality, it would also introduce additional overdiagnosis and overtreatment. It is therefore important to not only look at mortality reduction, but also take quality of life (QALYs) into account given the high side effects of treatment.\textsuperscript{11} In the Netherlands, there is no screening program for prostate cancer as it is expected that it would lead to drastic overtreatment.\textsuperscript{12} Additionally, due to the risk of overdiagnosis, general practitioners are advised to discuss the downsides of performing a PSA test with patients.\textsuperscript{13} Although there is no official PSA screening program in The Netherlands, a PSA test is often still the first step in the diagnostic pathway of prostate cancer. The PSA test is commonly combined with a digital rectal examination (DRE) by a urologist. When there is a further confirmed suspicion of prostate cancer, traditionally, the next step would be to take systematic transrectal ultrasound-guided (TRUS) biopsies. During this procedure, eight to twelve biopsies are taken from different regions in the prostate guided by an ultrasound probe.\textsuperscript{14} It has been shown that taking twelve biopsies increases the cancer detection rate, but the TRUS biopsies still have limited sensitivity for prostate cancer.\textsuperscript{15,16} With advances in imaging techniques, it recently became a new standard in the Netherlands to perform an MRI scan before a biopsy procedure to reduce the number of unnecessary biopsies.
1.2 Prostate cancer

Figure 1.4: Example of H&E-stained prostate tissue displayed using various degrees of magnification.

of invasive biopsy procedures. Still, to fully confirm the presence of prostate cancer, histological assessment of the tissue is required. If the MRI is positive, it is advised to perform an MRI-guided biopsy procedure, with optionally additional systematic biopsies. The combination of MRI-guided and systematic biopsies has been shown to be more specific. If the MRI was negative, in consultation with the patient, a further biopsy procedure can be omitted.

1.2.3 Histology

After a prostate biopsy is taken, the sample is sent for histopathological assessment by a pathologist. Histopathology is the medical science that studies the microscopic structures of tissue (histology) to understand human diseases (pathology). In clinical diagnostics, pathologists analyze samples of human tissue using bright-field microscopy in order to determine a diagnosis for the patient.

After the biopsy procedure, whether TRUS-, MRI-guided or systematic, the specimen cannot be directly examined by pathologists. The tissue is first fixated to prevent decay by immersing the sample in formalin; this ensures that the sample can be preserved for extended periods. After fixation, any remaining water is removed, and the sample is embedded in paraffin (FFPE, formalin-fixed paraffin-embedded). From this embedded block, very thin sections can be cut using a microtome—typically only
a few micrometers thick. After cutting, the sections are mounted on glass slides for easier handling and to enable microscopic inspection. When mounted, a staining procedure is applied to highlight specific parts of the tissue. The most common staining procedure worldwide stains tissue using hematoxylin and eosin (H&E), with hematoxylin staining nuclear DNA in blue and eosin staining cytoplasmic and tissue proteins in shades of pink (Figure 1.4). Between labs, or even within the same lab, the exact color characteristics can vary drastically due to (small) differences in the procedures and environmental influences. Besides H&E, other staining methods can highlight features of the tissue that are otherwise hidden or difficult to assess using H&E alone. Immunohistochemical (IHC) staining allows detecting specific proteins expressed by the cells contained in the tissue section. Although immunostaining highlights specific tissue proteins, assessing the overall architectural patterns of the tissue is difficult compared to H&E stained slides. Immunohistochemistry is also typically more expensive and resource-intensive as compared to the regular H&E staining. For prostate cancer, the tissue samples are typically assessed using H&E, while specific stains are used when additional evidence for a diagnosis is required.

### 1.3 Gleason grading system

Pathologists assess H&E stained prostate specimens following the Gleason grading system, which is the most important tissue-based prognostic marker for prostate cancer patients. Pathologists assess the architectural growth patterns of the tumor and assign it a number between 1 and 5, with increasing numbers corresponding to a decrease in histological differentiation and, typically, worse prognosis (Figure 1.3). While immunohistochemistry can be used to determine whether a case contains cancer, there is no specific stain that highlights Gleason patterns. Due to updates in the grading system, patterns 1 and 2 are not or rarely reported anymore for biopsies. The detected patterns are then summarized in the **Gleason score** of a specimen. For biopsies, the Gleason score is the sum of the most occurring pattern and the highest of the remaining patterns (e.g., 3+4=7). The second pattern should account for at least 5% of the tissue or be of a higher grade for it to be included in the reported score. For example, a biopsy with a majority pattern 4 and a second pattern 3 with a prevalence of 3% is assigned a score of 4+4=8. For the Gleason score of prostatectomies, the sum of the most common and the second most common pattern is used instead. Theoretically, the lowest Gleason score is 1+1=2, but in clinical practice the lowest reported Gleason score for a patient can be 3+3=6 and the highest 5+5=10.
1.3 Gleason grading system

Figure 1.5: Overview of Gleason’s patterns. Due to changes in the grading system, patterns 1 and 2 are rarely reported anymore. Image source: Wikimedia Commons/National Institutes of Health.

Even though the grading system is clearly defined, prostate cancer is a heterogeneous disease with a wide range of tissue patterns, which complicates grading. The assessment of the tumor grade is a subjective and challenging process and therefore suffers from significant inter- and intraobserver variability.\textsuperscript{20,21}

A 5-tier grading system was introduced to improve the reporting of Gleason grading, which directly assigns Gleason scores to one of five prognostic groups (Figure 1.6).\textsuperscript{19,22,23} Throughout the literature, the new grading system is referred to using different names, including (ISUP) grade group, Gleason grade groups, and ISUP grade. In this thesis, we use the general term grade group to refer to the new grading system. The groups, with the descriptions derived from Epstein et al.,\textsuperscript{19} are defined as follows:

**Grade Group 1:** *Gleason score* $3+3 = 6$, Only individual discrete well-formed glands.

**Grade group 2:** *Gleason score* $3 + 4 = 7$, Predominantly well-formed glands with lesser component of poorly-formed or fused or cribriform glands.

**Grade group 3:** *Gleason score* $4 + 3 = 7$, Predominantly poorly-formed, fused or cribriform glands with lesser component of well-formed glands.

**Grade group 4:** *Gleason scores* $4 + 4 = 8$, $3 + 5 = 8$, $5 + 3 = 8$, Either (1) only poorly-formed/fused/cribriform glands; (2) predominantly well-formed glands and lesser component lacking glands; or (3) predominantly lacking glands and a lesser component of well-formed glands.
The introduction of grade groups had two clear advantages. First, before the new system, the lowest reported score was a Gleason score $3 + 3 = 6$, and a score of 6 out of 10 does not communicate the relatively positive prognosis well, especially to patients. Secondly, as there is a clear prognostic difference between Gleasons score $3 + 4 = 7$ and $4 + 3 = 7$, the new grade group system explicitly communicates this by assigning these scores to different grade groups. While grade groups’ introduction showed clinical value and increased interpretability of the tumor grade for patients,\textsuperscript{24} there is also criticism on the simplification\textsuperscript{25}, and it has not addressed the inter- and intraobserver variability.\textsuperscript{26,27}

Unfortunately, a small deviation in tumor grade can already have considerable implications for individual patients. For example, whereas the recommended treatment for grade group 1 is active surveillance, a change to group 2 could result in advice for radical therapy (e.g., surgery, radiotherapy).\textsuperscript{28} Whether a patient would receive surgery can depend on which pathologist handles their case. Research from The Netherlands showed that, when biopsies of patients were re-evaluated, 21% of cases were assigned a different grade, and in 43% of those cases, the original diagnosis was upgraded.\textsuperscript{29} Such re-evaluation is not common because of associated high costs and is only done in 3% of the cases, often when patients are referred to another hospital.

Specialized uropathologists show higher concordance rates in Gleason grading, but such expertise is not always available.\textsuperscript{30} This is furthermore complicated by a global shortage of pathologists we are increasingly witnessing,\textsuperscript{31} limiting access to accurate diagnostics for many patients. In Western countries, such as the UK and the USA, there is at least one pathologist per 50,000 patients, though the workload is high,
1.4 Digital & computational pathology

While the traditional pathologists’ workspace was, until recently, based on a microscope, this is slowly transitioning to diagnosis using a computer screen: Digital Pathology. This digital transition is made possible by the introduction of whole-slide scanners that are able to scan individual glass slides at high resolution, resulting in whole-slide images (WSI).

Histological assessment of tissue requires a pathologist to view the tissue at high magnification. To allow for digital assessment, a scanner takes a gigapixel image of the tissue, typically scanned at a resolution of $0.25 \sim 0.50\mu m$ per pixel, corresponding to respectively a 40x and 20x objective of a traditional microscope. The pixel spacing of a scanned slide is crucial metadata as it defines how much area an individual pixel represents of the original tissue. A typical WSI can easily run up to 200,000 by 100,000 pixels, roughly corresponding to the same area as 1600 photos taken with a 12-megapixel camera. This considerable size introduces significant challenges in storing and management of these images.

WSIs are typically examined using specialized viewer software, which allows pathologists to efficiently navigate the image, zoom in on specific regions, and add annotations. One of those applications is the open-source ASAP viewer$^{32}$ built on top of OpenSlide$^{33}$ which has been used as the main viewer during the conduct of the research in this thesis.
Detect regions of interest, e.g. prostate cancer.

Detect & label regions of interest, e.g. Gleason's patterns.

Give case or patient-level diagnosis, e.g. grade groups or staging.

Figure 1.8: Abstract representation of different methods of AI-assistance. Automated systems can give feedback on multiple abstraction levels. For prostate cancer, this can range from tumor detection (left) to grading of regions using the Gleason grading system (center) and determining the biopsy or patient-level diagnosis (right).

Digital pathology can have significant improvements on the histopathology workflow. Benefits of digital pathology include remote diagnostics, streamlined external consultation, and the ease of access to archived cases. Where traditionally fragile glass slides had to be shipped physically, digital pathology moves this transfer to the digital domain. Of course, this digital transition co-exists with significant digital infrastructure requirements and additional measures for privacy and data integrity.

The digital transition within pathology also allowed for a new research field to emerge: computational pathology. With histological slides now digitized, it is possible to annotate patterns in the tissue specimens digitally. This annotated data can then be used to develop computer models to assist in diagnosis.

1.5 Computer-aided diagnosis

The digitalization of various medical domains, including pathology, introduced the possibility of assisting medical professionals with (intelligent) algorithms that assist in or automate part of the diagnostic process. On an abstract level, these algorithms take patient data as input and transform this data into some outcome, potentially relevant to the task. Formalized, we can define such algorithms as a mapping from some input $x$ to an output $y$:

$$ y = f(x) $$

This mapping can be simple, e.g., an algorithm that automatically computes the body-mass index (BMI) of a patient, which takes body weight and length as input:
While this example might sound trivial and would not be classified as an intelligent decision support tool, it highlights these algorithms’ basic structure. The same input-output-mapping definition holds for more complex systems, though these will have more elaborate inputs and outputs. The mapping function \( f() \) can be pre-defined, as in the BMI example, or learned. This thesis focuses on algorithms that take medical images as input and make predictions based on those images.

The assistance by algorithms can be performed on multiple abstraction levels, based on the output of the algorithm (Figure 1.8). At the lowest level, an algorithm could assist in the detection or quantification of certain structures. These are often referred to as computer-aided detection (CADe) or computer-aided quantification (CADq) systems. The task of such systems is to detect certain structures of interest in a medical image. Examples are the detection of lung nodules in CT,\(^\text{36}\) the detection of lymph node metastases,\(^\text{37}\) or the detection of prostate cancer in biopsies.\(^\text{38}\) Often, this detection is performed at a local level, so the input to the algorithm can also be limited, e.g., an algorithm could only require a small region around a lung nodule to detect it.

One level further concerns algorithms that focus on diagnosis, i.e., computer-aided diagnosis (CADx). These systems often are classification systems. Besides detecting specific objects of interest, these algorithms also need to classify the detected objects into pre-defined groups or classes. Examples of such classification systems are the classification of lymph node metastases in either micro, macro, or isolated tumor cells,\(^\text{39}\) or the assignment of Gleason patterns to individual prostate glands. Depending on the task, such systems could require a more extensive input. E.g., for detecting Gleason patterns, the surrounding area of a gland is required to make an accurate assessment of the morphology.

On the highest abstraction level, systems give feedback or output on a case or patient level. Algorithms that perform this task, especially in medical imaging, are often more complicated, as such systems take low-level input and transform this to high-level predictions. An example for prostate cancer could be the biopsy-level diagnosis, which summarizes the detected tissue patterns in a single grade group. On a patient-level, usually, information from multiple images is combined, e.g., the staging of a set of lymphnodes\(^\text{39}\) or the prognosis for an individual patient.

The names for the different abstraction levels of intelligent algorithms are often used interchangeably. In this thesis, we will use the general term computer-aided diagnosis (CAD), deep learning system, or algorithm as an overarching term.
Figure 1.9: Simplified representation of a single-layer perceptron, the most basic version of an artificial neural network. The neuron takes the weighted sum of its inputs and then applies an activation function.

1.6 Deep learning

Traditionally, CAD systems for medical imaging were designed by hand with so-called “hand-crafted features." Based on domain knowledge, researchers would engineer quantitative features to extract information from an image and create a representation on which machine learning models could be trained. Such algorithms can be seen as a two-step approach. First, the input is transformed using some pre-defined feature extractor function $h()$, after which a machine learning model $f()$ is trained to map these features to the desired output $y$:

$$y = f(h(x))$$  \hspace{1cm} (1.3)

This two-step approach was mostly due to machine learning techniques’ inability to learn from the raw data directly. Though, a minor benefit of this approach is that it is an easy way to introduce expert knowledge into the algorithm. However, this comes at the cost of having to pre-define the features instead of learning these directly from the data itself.

Currently, the field has almost entirely transitioned to using feature or representation learning. Instead of hand-crafting the features, the features are learned automatically from the raw data. At the basis is a technique that has been around for years: artificial neural networks.\textsuperscript{40} Such networks consist of connected artificial neurons (Figure 1.9), which through connections between the neurons form a network. Each neuron is a simple mapping function, but these networks can model complex functions as a whole. When there are many layers of neurons in a network, we speak of a deep neural network and the more general term: deep learning.
Historically, artificial neural networks, specifically multilayer perceptrons, consisted of fully connected layers in which each neuron was connected to all neurons in the previous layer. This architecture made those networks prone to overfitting and deep networks infeasible due to the memory requirements.

In recent years, advances have been made in the training of deep neural networks, further boosted due to the developments in hardware technology, specifically fast graphics processing units (GPUs). Due to these developments, the field underwent a deep learning revolution, making deep learning-based approaches the de-facto standard for many image analysis tasks. A famous example is a study by Krizhevsky et al. on the ImageNet classification task, whose deep learning approach outperformed the previous non-deep learning approaches by a wide margin. The available of many high-quality Open Source tools for deep learning has also stimulated further development.

One specific subtype of neural networks is crucial for computer vision: convolutional neural networks (CNN). Instead of the fully connected networks, CNNs instead use small filters that are slid over the input image at each valid position to compute the output. The benefit of these filters is that they are invariant to translations, as each filter is applied to every possible valid input region of the image. This approach makes CNNs highly effective for image analysis. Deeper into the network, each layer will respond to more complex features in the images, which is why CNNs and filters are often compared to the visual cortex (Figure 1.10).

A CNN consists of multiple building blocks that stacked together form the final network. Layers are connected in the form of a directed acyclic graph, where each layer represents a node to form the CNN. In its basic form, each layer directly follows the
previous layer. More complex architectures use, e.g., skip connections or concatenation operations to introduce additional parallel paths inside the network. While many variants and subvariants of these building blocks exist, there are a few of these that are often used:

**Convolutional layers** are at the core of a CNN and apply the filters to the input image. The number of filters per layer can differ and is often increased for deeper layers. Filters in earlier layers often react to basic features, while deeper layers typically respond to more complex structures of the input domain (Figure 1.10).\(^1\)

**Activation layers** apply a function to their input. From the biological perspective, this layer determines whether a neuron fires. These functions are often nonlinear.

**Pooling layers** perform a summary function, e.g., average or maximum, on a window of the input. Pooling layers downsample the input, resulting in a smaller input for subsequent layers and a larger field of view. Upsampling layers can reverse pooling layers, which is a key element of the popular U-Net architecture.\(^2\)

**Fully connected layers** are often used in the last part of a network to perform the final classification. At this point in the network, the previous layers have already extracted relevant features from the data, and the role of the fully connected layers is to incorporate this information in a final prediction. A convolutional layer can also emulate a fully connected layer by using a kernel size of 1. Doing this creates a fully convolutional neural network that can easily scale to larger input sizes.\(^3\)

### 1.7 Challenges for computational pathology

The field of computational pathology is characterized by specific challenges that distinguish it from other (medical) imaging domains. Here we discuss the three challenges most relevant to this thesis: the gigapixel size of the slides, artifacts, and variations in the images caused by different scanners and labs.

The first challenge is the size of the whole-slide images. Glass slides are typically scanned at \(40\times\) or \(20\times\) magnification, resulting in gigapixel-sized images with dimensions of \(200,000 \times 100,000\) pixels being no exception. Because of this size, exhaustively annotating all the tissue by hand is often infeasible or cost-ineffective for large datasets. Technically, annotating these slides by hand is still possible, but limited resources require innovative approaches to acquire enough training data.
1.7 Challenges for computational pathology

Besides an enormous annotation burden, the image’s size introduces technical problems during the training of neural networks. These gigapixel images do not fit a modern GPU’s memory at full resolution, limiting options for training deep learning models on the full image. While it is possible to downscale the whole-slide image to a resolution that would fit the GPU, this removes the necessary detail for many diagnostic tasks.

Instead of using the entire image, deep learning systems for pathology are typically trained in a so-called “patch-based” manner (Figure 1.11). The full image is split into small tiles (patches) that do fit into GPU memory and are then used for training. The same approach is utilized when a deep learning system needs to classify an unseen case (i.e., inference). The final prediction can be formed by processing the case tile-by-tile and stitching the output together.

In its typical form, patch-based training requires that individual patches are independent of each other, as the field of view of the deep learning system is limited to a single patch’s dimensions. For many tasks, this is not a limitation as predictions can be made using a relatively small field of view. However, if more global information is required, e.g., structure’s position within the tissue, solely relying on patch-based training will not work. A second limitation of patch-based training is that each patch requires a label. Due to large images in pathology, a slide-level label is not always represented by each patch. For example, a biopsy containing a tumor could have a slide-level label of ‘malignant,’ but not every region in that biopsy has to contain cancer.

A second challenge for automated analysis of histological slides is the presence of artifacts (Figure 1.12). The road from patient to a digitized slide of a patient’s tissue...
sample is long, and in each step of the process, artifacts can be introduced.\textsuperscript{54} An overview of common artifacts:

**Out-of-focus regions** are caused by incorrect focus points of digital pathology scanners. Typically scanners start with a low-resolution scan to determine focus points. If these focus points are set incorrectly, parts of the tissue will be rendered out-of-focus, which hinders the tissue’s assessment (Figure 1.12a).

**Tissue folds** occur during handling of the tissue sections. A fold distorts the tissue and can be recognized by a typical darker well-demarcated line (Figure 1.12b). Sometimes a tissue fold also introduces blurring artifacts if the scanner selected an incorrect focus point due to the tissue fold being thicker than the surrounding area.

**Ink** is often used to aid the identification of tissue samples. For example, biopsies can be inked with distinct colors before cutting and then placed on the same glass slide. The presence of ink helps to identify the biopsies in the final scan. Unfortunately, (excessive use of) ink can also obscure some of the tissue and result in misclassifications by deep learning systems (Figure 1.12c).

**Dust, air bubbles** or even eyelashes of the technician can end up under the coverslip. After scanning, these artifacts will be clearly visible in the image (Figure 1.12d and 1.12e).

**Pen markings** are commonly present in archived glass slides. In the era of the microscope, pen markings were an easy way of annotating or outlining tissue regions for future reference. When scanning large sets of archived slides, it can be expected that these pen markings are present if the slides are not cleaned first (Figure 1.12f). Due to the introduction of digital pathology, the presence of pen markings will decrease over time.

The last challenge regards scanner and stain variations. Image characteristics can vary wildly based on the lab that stained the tissue and the scanner that digitized the glass slide (Figure 1.13).\textsuperscript{56} These variations can be challenging for deep learning methods if the algorithm is purely trained on data from a single center and needs to be applied to data from a different center.

Some of the variations can be circumvented through simple measures. For example, discrepancies in the pixel spacing between scanners can often be reduced by resizing the image, e.g., when an image is scanned at a pixel resolution of 0.26\(\mu m\), and the network was trained on data scanned with a resolution of 0.24\(\mu m\). Stain normalization or transformation techniques can also be used to reduce stain variations.\textsuperscript{57–61}
Figure 1.12: Examples of common artifacts found in histological slides: (a) out-of-focus regions, (b) tissue folds, (c) ink, (d) dust, (e) air bubbles, and (f) marker.

Figure 1.13: Examples of breast tissue samples stained with H&E from different centers.\textsuperscript{55} Even though the same stain is used, because the tissue is stained in different centers, there is significant stain variance.

Alternatively, algorithms can be trained using data augmentation that tries to simulate stain variations and scanner characteristics.\textsuperscript{62} In this approach, instead of trying to remove variation in the new data, additional variation is introduced during the training phase.
1.8 Aims and objectives

The histological grading of prostate biopsies is a crucial element in the diagnostic pathway of prostate cancer. The known high inter- and intraobserver variability show potential and a need for assisting pathologists in this task. Furthermore, a global shortage of pathologists stresses the demand for reproducible, more efficient, and easily accessible diagnostic solutions.

This thesis’s primary aim was to investigate and design an AI-based system to detect and grade prostate cancer in biopsies. A second aim was to evaluate the potential clinical merits of AI-assisted grading when such systems are embedded in the pathologist’s workflow. To this extent, the following objectives were undertaken as part of this thesis:

1. The development of an automated system that can distinguish epithelial tissue from other tissue types within H&E stained prostate specimens (Chapter 2);
2. The development and validation of an automated system for grading prostate biopsies using the Gleason grading system (Chapter 3);
3. A multi-center independent evaluation of state-of-the-art algorithms for automated Gleason grading sourced through a large-scale medical AI competition (Chapter 4);
4. The investigation of the potential merits of AI-assisted grading of prostate cancer through an observer study (Chapter 5).

1.9 Thesis outline

The remainder of this thesis is organized in the following chapters (Figure 1.14):

Chapter 2 Detecting epithelial tissue in H&E can be seen as a prerequisite for the detection and grading of prostate cancer. Manual annotation of epithelial tissue is complicated by the heterogeneous character of epithelial tissue within the prostate, ranging from intact glands to strands of individual cells. This chapter describes a method to train a deep learning system to segment epithelial tissue using immunohistochemistry as the reference standard, alleviating the need for manual annotations.

Chapter 3 In order to assist pathologists when grading prostate cancer, deep learning methods need to achieve pathologist-level performance. This chapter investigates the development and validation of a deep learning system for the
Figure 1.14: Overview of main thesis chapters. This thesis describes the full process from the development of AI for Gleason grading to the evaluation of the merits of such systems.

diagnosis of prostate cancer in biopsies. For the development of the deep learning system, a completely new cohort of prostate cancer patients was collected. Furthermore, the deep learning system is compared to pathologists on several clinically relevant decision thresholds.

Chapter 4 Clinical implementations of automated Gleason grading requires algorithms that perform well in a multi-center setting. In this chapter, several state-of-the-art algorithms are investigated and independently validated using data from multiple centers through an international competition: the PANDA challenge.

Chapter 5 When AI algorithms have shown to achieve pathologist-level performance at diagnosing prostate cancer, it remains to be investigated how this could benefit clinical diagnostics. This chapter investigates the effect of AI assistance on Gleason grading by pathologists.

Chapter 6 The final chapter summarizes the key contributions of this research and discusses the implications and future work.
IHC-based epithelium segmentation

Authors: Wouter Bulten, Péter Bándi, Jeffrey Hoven, Rob van de Loo, Johannes Lotz, Nick Weiss, Jeroen van der Laak, Bram van Ginneken, Christina Hulsbergen-van de Kaa and Geert Litjens

Original title: Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard

Published in: Scientific Reports (Volume: 9, number: 864, January 2019)

DOI URL: doi.org/10.1038/s41598-018-37257-4
Abstract

Given the importance of gland morphology in grading prostate cancer (PCa), automatically differentiating between epithelium and other tissues is an important prerequisite for the development of automated methods for detecting PCa.

We propose a new deep learning method to segment epithelial tissue in digitized hematoxylin and eosin (H&E) stained prostatectomy slides using immunohistochemistry (IHC) as the reference standard.

We used IHC to create a precise and objective ground truth compared to manual outlining on H&E slides, especially in areas with high-grade PCa. 102 tissue sections were stained with H&E and subsequently restained with P63 and CK8/18 IHC markers to highlight epithelial structures. Afterwards each pair was co-registered.

First, we trained a U-Net to segment epithelial structures in IHC using a subset of the IHC slides that were preprocessed with color deconvolution. Second, this network was applied to the remaining slides to create the reference standard used to train a second U-Net on H&E.

Our system accurately segmented both intact glands and individual tumour epithelial cells. The generalisation capacity of our system is shown using an independent external dataset from a different centre.

We envision this segmentation as the first part of a fully automated prostate cancer grading pipeline.
2.1 Introduction

With 1.1 million new diagnoses every year, prostate cancer (PCa) is the most common cancer in men in developed countries.\textsuperscript{63} PCa develops from genetically damaged glandular epithelium, resulting in altered cellular proliferation patterns. In the case of high-grade tumors, the glandular structure is eventually lost and strands of (individual) cells can be observed instead.\textsuperscript{64}

The histological grade in PCa is formally defined in the Gleason grading system,\textsuperscript{65} and is a powerful prognostic marker. It is determined by pathologists on hematoxylin and eosin (H&E) stained tissue specimens. The grade is based on the architectural growth patterns of the tumor which are assigned a number between 1 and 5, with increasing numbers corresponding to a decrease in histological differentiation, and, typically, worse prognosis.\textsuperscript{18}

The identification and grading of prostate cancer can be time-consuming and tedious for pathologists, as all individual cancer foci within a surgical specimen or biopsy have to be analyzed. This is compounded by the fact that prostate cancer is generally a multi-focal disease and that surgical specimens can consist of anywhere between 8 to 15 sections. Although nowadays, thanks to the advent of whole-slide scanning systems, pathologists can perform their diagnoses on a computer screen instead of using a microscope, this has not directly helped them to perform more efficient or accurate diagnostics. However, computer-aided diagnostic tools based on deep learning and convolutional neural networks have shown promise in improving the accuracy and efficiency of histopathological diagnosis.\textsuperscript{38}

Deep learning methods that try to detect or grade cancer from scanned tissue slides are typically trained using a set of annotated regions as the reference standard. As these algorithms learn from training data, the quality of the output is directly linked to the quality of the training samples. Ideally, training samples for detecting and grading PCa consist of individually outlined glands. However, outlining PCa requires extensive expert knowledge due to the large differences between and within Gleason grades. In addition, annotating individual cells of high-grade PCa is practically infeasible due to the mixture of glandular, stromal, and inflammatory components. Therefore, tumor annotations made by pathologists are often coarse and contain large amounts of non-relevant tissue, which adds noise to the reference standard and, subsequently, limits the potential of deep learning methods.

We propose a method to automatically improve the detail of PCa annotations by pathologists by dividing digitized tissue into relevant and non-relevant tissue on a pixel-by-pixel basis, in this case epithelial versus other tissues. Such a system can help improve the detail of coarse cancer or grade annotations but can also be use-
ful by itself in highlighting areas containing epithelial cells as regions of interest for pathologists.

To train our system, we employed a novel two-step approach (Figure 2.1). First, we trained a convolutional network to segment epithelium in immunohistochemically (IHC) stained tissue sections applying an epithelial marker. By applying color deconvolution and subsequent recognition of positively stained pixels, we were able to have ample training data while obviating the cumbersome and imprecise process of manually annotating epithelial regions.\cite{66,67} Registration was used to map the network’s output to the H&E version of the specimens, which were subsequently used as training input for our final model. Our automated segmentation is not only useful as a tool for pathologists, we particularly envision this segmentation as being the first part of a fully automated prostate cancer detection and grading pipeline.

### 2.2 Related work

Existing research on segmenting epithelial tissue has shown promise in PCa specimens. Gertych et al.\cite{68} used a support vector machine to distinguish between stroma and epithelial glands and applied this to a dataset of 20 patients containing specimens of Gleason grade 3 and 4. Hand-crafted features, based on the intensity and spatial relationship of pixels, were derived from H&E specimens that had been preprocessed using color deconvolution. Naik et al.\cite{69} employ Bayesian classifiers to segment glands, relying on the presence of lumen in the glands. The segmentation was applied to Gleason grade 3 and 4, and benign tissue samples; not on the less common but more aggressive pattern 5. Gleason grade 5 can express in the form of single-cell strands or nests, or solid sheets (with or without central necrosis) of malignant cells with no or minimal lumen formation; obviously, this could hinder a segmentation method that relies on the presence of lumina. Singh et al.\cite{70} employed a multi-step approach based on logistic regression to segment epithelium, distinguishing between glands, lumen, peri-acinar retraction clefting and stroma. Both Gertych et al.\cite{68} and Naik et al.\cite{69} used the segmentation results as a first step towards automated Gleason grading.

Advances in deep learning have resulted in new methods for performing segmentation. Deep learning methods generally outperform hand-crafted features on segmentation tasks in digital pathology, for example, on H&E and IHC stained breast and colon tissue specimens.\cite{71} On the dataset from Gertych et al.,\cite{68} Li et al.\cite{72} show a clear performance increase when using deep learning models to segment PCa in comparison to classical machine learning methods. Deep learning methods also show good performances on segmenting glands, for example in colorectal tissue.\cite{73}
2.2 Related work

1) Specimens are stained with CK8/18 and P63 to mark epithelial tissue and basal cell layer. 2) Color deconvolution is applied. Only the channel representing the epithelial tissue is used, the rest is discarded. 3) Artifacts are present due to imperfections in the staining and color deconvolution method.

4) Artifacts are removed manually in selected regions. 5) A 5-layer deep U-Net is trained on the corrected IHC regions. Areas with artifacts are sampled more. 6) The IHC network produces precise segmentation masks given an IHC slide, independent of the color deconvolution.

7) Slide pairs are registered on cell-level due to the use of restained slides and non-linear patch based registration.

8) The trained IHC network is applied to each IHC slide. The network output is used as the training mask for the H&E network. 9) A 6-layer deep U-Net is trained on H&E and the masks generated by the IHC network. The trained H&E network segments epithelial tissue on H&E.

Figure 2.1: Overview of methodology. We first train a network (1) on a subset of our IHC training data. The segmentations produced by this first network are then transferred to H&E and used to train the final network (2).
Table 2.1: Overview of case grading from original pathologist’s report on section level (using grade group) and on individual grade. Note that multiple grades can occur within a single slide.

<table>
<thead>
<tr>
<th>Set</th>
<th># slides</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train set</td>
<td>62</td>
<td>24</td>
<td>10</td>
<td>11</td>
<td>3</td>
<td>14</td>
<td>12</td>
<td>44</td>
<td>40</td>
<td>22</td>
</tr>
<tr>
<td>Test set</td>
<td>40</td>
<td>15</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>9</td>
<td>12</td>
<td>26</td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>102</td>
<td>39</td>
<td>16</td>
<td>18</td>
<td>6</td>
<td>23</td>
<td>24</td>
<td>70</td>
<td>64</td>
<td>34</td>
</tr>
</tbody>
</table>

Previously, we performed a pilot study on epithelium segmentation comparing U-Net versus regular fully convolutional networks using 30 radical prostatectomy slides and a small, manually annotated, test set. We achieved the best segmentation performance using a 4-layer-deep U-Net, but found that the performance of our network capped due to errors in the reference standard. Moreover, a low number of samples, in particular few high-grade PCa specimens, limits the applicability to daily practice.

Most of the existing studies on epithelium segmentation in prostate suffer from small datasets or focus on a subset of the occurring grades. In this paper, we did not exclude any Gleason grades or gland morphology.

2.3 Materials

We selected a cohort of 102 patients who underwent a radical prostatectomy at the Radboud university medical center (Radboudumc) between 2006 and 2011 (IRB number 2016-2275). Patients who received adjuvant therapy before surgery were excluded. From each prostatectomy, we selected one formalin fixated paraffin-embedded tissue block based on the Gleason grades reported in the original pathologist’s report. Based on the reported grades, we determined the Gleason grade group for each block (Table 2.1). As a tissue block can contain multiple grades, we also reported the individual occurrences of each grade. Of all tissue blocks, 24% contained a region with grade 2, 69% with grade 3, 63% with grade 4, and 33% with grade 5. Due to selective oversampling, the incidence of high-grade tumors (grades 4 and 5) is relatively higher than in clinical practice. This oversampling allows us to explicitly investigate the performance of deep learning-based epithelium segmentation algorithms on high-grade PCa, in which such segmentation is most challenging.
2.3 Materials

Figure 2.2: Dataset examples (first and third row H&E, corresponding IHC on the second and fourth row). Our restaining procedure (instead of using consecutive slides) results in perfectly matching slides. The first five examples show benign epithelium, and the last five show various grades of PCa. In the IHC examples, all epithelial tissue is marked in brown, the basal cell layer in dark red (only present in the benign examples). Between cases, the intensity of the stain can differ substantially.

From each block a new section was cut, stained with H&E and scanned using a 3DHistech Pannoramic Flash II 250 scanner. After scanning, the tissue was destained, restained using immunohistochemistry, and scanned again. All slides were scanned at 20x magnification (pixel resolution 0.24 $\mu m$).

We used two markers for the immunohistochemistry: CK8/18 (using DAB) to mark all glandular epithelial tissue (benign and malignant), and P63 (using NovaRED) for the basal cell layer, which is normally present in benign glands but not in malignant glands. This staining procedure results in a slide where all relevant tissue is highlighted, providing us with a clear ground truth (see Figure 2.2 for examples). Staining the basal cell layer using a different color makes it easier to spot tumor regions in IHC and can facilitate grading of the tissue on H&E. Restaining, instead of making
consecutive slides, results in a H&E and IHC whole-slide image (WSI) pair for each patient that contains the same tissue. Although the slide pairs were made from the same glass slide, minor alignment errors and tissue deformations were still present due to the restaining procedure.

The 102 scanned slide pairs were split into two sets: a training set (62) and a test set (40). The slides were distributed over the sets at random while stratifying for Gleason grade group (Figure 2.3). The test set was used as a hold-out set and not used during training or model optimisation.

2.3.1 Hold-out test set

For each IHC slide in the test set, a trained non-expert divided each WSI in four sections: two containing tumor and two containing only benign epithelium. From each of these four regions, we extracted an area of 2500 × 2500 pixels randomly at 10x magnification. If there was either no tumor or benign region available, an additional region from the other category was selected. This method resulted in 160 regions. The tumor regions were individually graded by an experienced pathologist (C.H.-v.d.K.) with subspecialty uropathology, without using the original patient’s record. We recorded the primary, secondary and tertiary (if present) grade for each region (Figure 2.4). The reported grades were not necessarily identical to those from the patient records; the selected regions contained a subset of the slide and were extracted from a newly cut section. The Gleason grade group was based on the ISUP scoring system for biopsies (most prevalent plus highest grade).

2.3.2 External test set

Gertych et al. made their dataset available to use for external validation. This set consists of 224 1500 × 1500 pixels tiles sampled from 20 digitized WSIs (pixel resolution 0.5 µm) of H&E prostatectomy specimens containing Gleason grades 3 and 4. The tiles were already annotated by two pathologists, and each pixel labeled as stroma, benign epithelium, Gleason 3, or Gleason 4. Glands were annotated as a whole, including the lumen. We combined the annotations of benign epithelium and the two PCa grades into a single epithelium class.
2.3 Materials

**Figure 2.3:** Distribution of Gleason grade groups for each case in our dataset as reported in the original pathologist's report (N=102).

**Figure 2.4:** Gleason grades of tumor regions in the hold-out test set (N=71). Showing individual occurrences (left, 1-3 per region) and grade groups on region level (right).
2.4 Methods

We took a two-step approach to train a system for segmentation of epithelial tissue on H&E histopathology. First, we circumvented the challenge of manually annotating tissue by generating precise training data using immunohistochemistry and training a network on IHC. Then we transferred the output of the first network to H&E and trained the final segmentation network. Our networks were built using Keras and Tensorflow.

2.4.1 Slide preparation

We applied a pre-trained tissue-background segmentation network to all slides in order to exclude areas not containing tissue from further analysis. Next, color deconvolution was applied to all IHC WSIs in our training set. The resulting P63/CK8-18 channel was then converted to a binary mask by thresholding. Small errors were removed automatically using binary closing and opening. The resulting masks were not perfect due to imperfections and intensity changes in the stain, scanning artifacts, and non-specific staining; e.g., corpora amylacea and debris inside the glands are regularly stained brown and are therefore present in the deconvolution mask (Figure 2.5a,c).

For the hold-out test set, three trained non-experts reviewed the sampled test regions and manually updated the color deconvolution mask, removing any artifacts or updating incorrectly labeled tissue.

2.4.2 Training a CNN on IHC

Due to time constraints, it was unfeasible to manually correct all individual color deconvolution masks to be used for training. Instead, we trained a deep convolutional network to perform the mapping from a P63/CK8-18 slide to a binary epithelium mask. We selected 25 slides from our training set to train this first network (20 for training, 5 for validation). On each slide, we outlined a tissue region covering roughly 50% of the WSI, after which three trained non-experts corrected the color deconvolution masks by hand. A total of 3493 annotations were made by the annotators on these 25 slides, an average of 140 annotations per slide. In terms of surface area, 2.3% of the tissue was given a different label by the annotators. On average, the annotators took 45 to 60 minutes to correct a slide.

We trained a five-level-deep U-Net on the selected regions to segment epithelial tissue in IHC slides. We followed the original U-Net model architecture but added additional skip connections within each layer block and used up-sampling operations.
2.4 Methods

![Figure 2.5](image)

**Figure 2.5:** Effect of stain artifacts on network predictions. In some cases, non-epithelial tissue is stained, e.g., structures inside the gland (a, corresponding H&E version shown in b), which are picked up by the color deconvolution algorithm (c). Due to a high frequency of these artifacts, a network trained on this uncorrected data will have a high number of false positives in its predictions (d). Training on manually corrected data instead results in a better segmentation (e). These errors transfer to the training of the H&E network. A network trained on the raw color deconvolution masks makes more mistakes in these artifact regions (f) than a network trained on the output of the corrected IHC network (g).

The network was trained using randomly sampled patches with a size of $512 \times 512$ (pixel resolution $0.48 \mu m$) and a batch size of 1. Regions with annotated artifacts and corpora amylacea were oversampled to lower the number of false positives. Adam optimisation was used with $\beta_1$ and $\beta_2$ set to 0.99, and a learning rate of 0.0005. The learning rate was halved after every five consecutive epochs without improvement on the validation set.

During training, we applied data augmentation to prevent overfitting and to improve the model’s generalization. The following augmentations were used: flipping, rotation, additive Gaussian noise, Gaussian blurring, and changes in saturation, contrast and brightness. After training, the model was applied to all IHC WSIs in our training set. A binary mask was created from each slide using the argmax of the network
output. We focused explicitly on color augmentations to overcome the large stain differences between the IHC slides.

For comparison, a second U-Net was trained on the non-corrected color deconvolution masks directly, without using any of the manual corrections. All hyperparameters and network structure were kept the same as in the original experiment to create a fair comparison.

2.4.3 Registration

The H&E slides were registered to the IHC slides using a nonlinear image registration method based on a method described previously. Since both slide images showed the same object with different stains, they were already approximately aligned. However, additional nonlinear deformations are caused by the chemical treatment during restaining and/or the slide scanning procedure and needed to be compensated for. Since different stains are used in both images, the colors of spatially corresponding structures do not match (Figure 2.2). We use the Normalised Gradient Fields (NGF) distance, which measures the alignment of image gradients, to account for the multimodality of the registration problem.

The registration pipeline consisted of: conversion of RGB images to gray-scale → para
drmetric (affine) registration → nonparametric registration (NGF distance measure, curvature regulariser) → patch-based registration (NGF, curvature). The method to merge the patches has been extended as follows: Instead of averaging the deformation patches, an optimisation problem is solved that balances data-fit and global deformation regularisation in the overlap region.

2.4.4 Training a CNN on H&E

The training masks generated by the IHC network matched the H&E slides as a result of the registration step; 50 were used for training and 12 for validation. We found that increasing the depth of the U-Net lowered the number of misclassified corpora amylacea on H&E. Therefore, for the H&E segmentation we trained a six-level-deep U-Net in comparison to the five-level-deep IHC network. To limit the parameter count caused by the added level, we lowered the number of filters for each level. The same extensions as used in the U-net for the IHC stained images were applied. The network was trained using patches with a size of $1024 \times 1024$ (pixel resolution $0.48 \mu m$) and a batch size of 1. Adam optimisation was used with $\beta_1$ and $\beta_2$ set to 0.99, and a learning rate of 0.0005. The learning rate was halved after every ten consecutive epochs without improvement on the validation set. The following data augmentations were
used: random scaling, flipping, rotation, additive Gaussian noise, Gaussian blurring, and changes in saturation, contrast, brightness and Haematoxylin-Eosin color space. Only the binary segmentation masks generated by the IHC network were available for training. We did not correct the masks manually. This meant that the sampling technique used for training the IHC network could not be applied to the H&E network. Instead, we sampled uniformly over the classes. To force the network to learn small areas of epithelium, e.g., in cases of Gleason 5, we weighted the loss of each pixel based on the class occurrence within a patch. As a result, even patches with only small individual tumor cells were picked up by the network due to a higher loss contribution.

To test the merit of the IHC network as input for our network, we also trained a U-Net on the raw color deconvolution masks. All hyperparameters and network structure were kept the same in both experiments.

2.4.5 Evaluation

The trained H&E network was applied to all WSIs of our hold-out set and evaluated within the randomly selected regions. No further post-processing was performed. The annotations of the external set were coarse and on gland-level (i.e., including the lumina) and did not match the output of our network. In accordance with the method used in the original paper, we removed the background from the color-normalized images of the external test set. Lumina (consisting of pixels that are classified as background pixels) were not used in computing the scores. We then fed the images to our trained H&E network. We did not optimize our network on this external set. As such, the results on the external test set can be considered a true estimate of the generalization capacity of our H&E network.

2.5 Results

We evaluated both the IHC and H&E networks on the regions from, respectively, the IHC and the H&E WSIs from the hold-out test set. The network output was compared with the ground truth: color deconvolution masks generated from the IHC slides with manual corrections. We report pixel-based accuracy, F1-score and Jaccard index using epithelium as the positive label (Table 2.2).
Figure 2.6: Zoomed-in examples (1000 × 1000 crop) of the hold-out test set: IHC version (left), ground truth (middle), and segmentation of the IHC network (right). Green pixels show true positive, red false positive, and blue false negative. The first example (a) shows an almost perfect segmentation. In regions where the stain is light or absent, the performance degrades (b).

2.5.1 Segmentation performance on IHC

The IHC network achieved an overall F1 score of 0.915. Given a minimum F1 score of 0.352 and a maximum of 0.980, the range of scores was high. Some regions of our test set suffered from an overall low stain quality or contained areas where the epithelium reacted less to the stain. We observed that a lower stain intensity resulted in a lower performance (Figure 2.6). As the H&E network was trained on the output of the IHC network we considered the IHC performance as an upper bound for the performance of the H&E network.

Using the corrected color deconvolution masks as training data resulted in an F1 score increase of 0.909 to 0.915 on our test set (Table 2.3). The network that was trained on the uncorrected data makes more mistakes in regions with stained non-epithelial tissue, e.g., corpora amylacea and other concretions inside the glands (Figure 2.5d).
2.5.2 Segmentation performance on H&E

The H&E network achieved an overall F1 score of 0.893. The score on benign tissue (F1 0.907) was slightly higher than on tumorous areas (F1 0.876). A decline in performance was observed in regions with higher Gleason grades. Regions with Gleason grade group 5 had an F1 score of 0.819. Several regions are displayed in Figures 2.7 and 2.8.

The score of the H&E network was comparable to that of the IHC network, showing that, given this training data, the network achieved almost optimal performance. Even more, the minimal performance of the H&E network was higher than the minimum of the IHC network (0.661 versus 0.352). Outliers that were present in the results of the IHC network were not present in the results of the H&E network.

Using the IHC network to generate training data, as opposed to the raw color de-convolution masks, resulted in an improved F1 score of 0.893 versus 0.878 for the uncorrected network (Table 2.3). Comparable to the IHC network, the uncorrected H&E network makes more mistakes in areas that are incorrectly targeted by the stain (Figure 2.5f).

2.5.3 Segmentation performance on external dataset

On the external set our network achieved an F1 score of 0.835 (Table 2.4, Figure 2.9). This is lower than on our hold-out test set, but within expectations due to the differences in staining and image resolution. With a Jaccard score of 0.735 we achieved a higher score than the original method, which had a Jaccard score of 0.595, and comparable to other deep learning methods that have been trained on this dataset.

2.6 Discussion

We developed a deep learning-based system that segments epithelial tissue in H&E-stained whole-slide prostatectomy images. Our system produces cell-level segmentations and is able to segment both intact glands as well as individual (tumor) epithelial cells. A common problem when training deep learning models for scanned histology sections is the absence of precise ground truth. We circumvented this problem by restaining our slides with an epithelial and basal cell layer marker. Using color de-convolution and a separately trained network, we were able to exhaustively annotate our complete training set with only a minimal amount of manual labor. This technique works especially well for annotating small instances of epithelium, e.g., cases
Figure 2.7: Zoomed-in example regions (1000 × 1000 crop) from the hold-out test set with H&E (left), ground truth (middle) and network segmentation (right). Green pixels show true positive, red false positive, and blue false negative.
2.6 Discussion

Figure 2.8: Two failure cases of the hold-out test set: a case of high-grade PCa (a) and a benign region (b) where debris inside the gland is segmented.

Figure 2.9: H&E network applied to cases from the external test set: (a) an example of a good segmentation, and (b) a case of undersegmentation.
of Gleason 5 PCa, that would most likely be missed by human annotators. Moreover, the use of specific markers renders our ground truth less subjective compared to manually produced annotations on H&E slides (even in inflamed or poorly differentiated areas). On an external test set, we see a drastic performance improvement compared to the original method, showing the generalization capacity of our network, even on images from an external center. When comparing to more recent deep learning methods on this dataset, we observe that our method performs as good. Of notice is that the methods we compare against were trained on the external test dataset (in cross-validation), whereas our network has never seen this data before.

In contrast to other previous work, we assess the performance of our algorithm across all Gleason grades, including the notoriously difficult Gleason grade 5. Although we do obtain the lowest score on this pattern (F1-score of 0.819), this score is still high especially given the poorly differentiated character of high Gleason grades, and the first benchmark on these grades. To allow others to compare their algorithms against ours, we have decided to release our test data and H&E WSIs publicly, including both the test and training slides. This dataset includes the 102 whole-slide H&E images used in this paper, all color deconvolution masks, and the manually corrected regions.

We train our IHC network on manually corrected regions which adds additional effort to the training procedure. These manual annotations result in a small increase in performance on our test set (F1 score 0.915) in comparison to training on non-corrected data (F1 score 0.909). Using the IHC network output to train the H&E network also improves its segmentation performance (F1 score 0.893 versus 0.878). While the numerical differences are small, using the corrected data is of importance in this particular dataset to lower the number of misclassifications that are caused by an aspecific stain (Figure 2.5) or in regions where the stain is absent. These consistent errors lower the applicability of the network in future systems. For other datasets, where stain artifacts are less prominent, training a network directly on the color deconvolution mask could be sufficient.

Our work also has some limitations. The method to establish the training labels is not perfect. The IHC network is only trained on a limited set of WSIs and is therefore not able to overcome all problems caused by stain variability and the presence of scan and tissue artifacts. Especially corpora amylacea or other debris inside glands, which are often stained by the epithelial marker, are a source of errors. Epithelium glands are also missed by the network when the stain is light or absent. Subsequently, misclassified areas on the IHC slides are transferred to the training data of the H&E network. Many of these errors are overcome by the H&E network due to the larger size of the H&E training set, which results in a much higher minimum performance with an F1-score of 0.661 vs. 0.352 for the IHC network.
The type of misclassifications is also influenced by the chosen magnification level. Low magnification is sufficient for segmenting intact glands, and could potentially help with lowering the number of artifacts as the network can learn high-level shapes of the tissue. However, segmenting individual epithelial cells, especially in the case of high-grade PCa, requires input patches with enough detail to be able to distinguish those cells from the surrounding stroma. We deliberately chose a high magnification level to improve the performance on high-grade PCa. In future work, it might be fruitful to investigate multi-scale approaches to tackle this issue.

We observe that the segmentation performance of our H&E network approaches that of the IHC network, which is used to generate the training reference for the H&E network. As a result, there is only a limited amount of improvement possible without further refining the training data. Annotating specific regions that are troublesome and retraining the IHC network on these regions could further boost the performance of the H&E network. However, one needs to consider that for some cells, it is simply impossible to assess their class using the H&E stain alone, especially in areas with active inflammation. As such, a perfect segmentation does not exist.

We see the development of an accurate epithelium segmentation network as the first part of a fully automated prostate cancer detection and grading pipeline. More specifically, the epithelium segmentation can be used to precisely outline potential cancer regions, and in combination with coarse tumor annotations, result in highly detailed annotations of PCa. We intend to leverage this to develop highly accurate PCa segmentation networks in the near future.

### 2.7 Data availability

The dataset generated during the current study is available in the Zenodo repository, [http://doi.org/10.5281/zenodo.1485967](http://doi.org/10.5281/zenodo.1485967).

### 2.8 Acknowledgements

This study was financed by a grant from the Dutch Cancer Society (KWF), grant number KUN 2015-7970. The authors would like to thank Milly van den Warenburg and Nikki Wissink for their help in making the manual annotations.
Table 2.2: Segmentation results on the hold-out test set.

<table>
<thead>
<tr>
<th>Regions</th>
<th>N</th>
<th>F1 score, mean (min, max)</th>
<th>Accuracy</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IHC network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All regions</td>
<td>160</td>
<td>0.915 ± 0.09 (0.352, 0.980)</td>
<td>0.952</td>
<td>0.854</td>
</tr>
<tr>
<td>Benign</td>
<td>89</td>
<td>0.944 ± 0.04 (0.712, 0.980)</td>
<td>0.980</td>
<td>0.897</td>
</tr>
<tr>
<td>Cancer</td>
<td>71</td>
<td>0.879 ± 0.11 (0.352, 0.974)</td>
<td>0.917</td>
<td>0.799</td>
</tr>
<tr>
<td><strong>H&amp;E network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All regions</td>
<td>160</td>
<td>0.893 ± 0.05 (0.661, 0.959)</td>
<td>0.940</td>
<td>0.811</td>
</tr>
<tr>
<td>Benign</td>
<td>89</td>
<td>0.907 ± 0.04 (0.780, 0.957)</td>
<td>0.966</td>
<td>0.832</td>
</tr>
<tr>
<td>Cancer</td>
<td>71</td>
<td>0.876 ± 0.05 (0.661, 0.959)</td>
<td>0.907</td>
<td>0.784</td>
</tr>
<tr>
<td>Grade group 1</td>
<td>32</td>
<td>0.884 ± 0.03 (0.808, 0.938)</td>
<td>0.921</td>
<td>0.793</td>
</tr>
<tr>
<td>Grade group 2</td>
<td>10</td>
<td>0.885 ± 0.03 (0.854, 0.927)</td>
<td>0.894</td>
<td>0.794</td>
</tr>
<tr>
<td>Grade group 3</td>
<td>5</td>
<td>0.893 ± 0.03 (0.833, 0.921)</td>
<td>0.912</td>
<td>0.809</td>
</tr>
<tr>
<td>Grade group 4</td>
<td>14</td>
<td>0.889 ± 0.06 (0.728, 0.959)</td>
<td>0.907</td>
<td>0.806</td>
</tr>
<tr>
<td>Grade group 5</td>
<td>10</td>
<td>0.819 ± 0.07 (0.661, 0.914)</td>
<td>0.874</td>
<td>0.699</td>
</tr>
</tbody>
</table>

Table 2.3: Comparison of segmentation performance of networks trained on the raw color deconvolution masks or using corrected training data.

<table>
<thead>
<tr>
<th>Training data</th>
<th>F1 score, mean (min, max)</th>
<th>Accuracy</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IHC network</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color deconvolution</td>
<td>0.909 ± 0.10 (0.312, 0.983)</td>
<td>0.951</td>
<td>0.844</td>
</tr>
<tr>
<td>Color deconvolution + corrections</td>
<td>0.915 ± 0.09 (0.352, 0.980)</td>
<td>0.952</td>
<td>0.854</td>
</tr>
<tr>
<td><strong>H&amp;E network</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color deconvolution</td>
<td>0.878 ± 0.06 (0.650, 0.954)</td>
<td>0.933</td>
<td>0.787</td>
</tr>
<tr>
<td>IHC network predictions</td>
<td>0.893 ± 0.05 (0.661, 0.959)</td>
<td>0.940</td>
<td>0.811</td>
</tr>
</tbody>
</table>

Table 2.4: Comparison of results on the external test set. Our method was not trained on this external set while the other methods have been trained using cross validation.

<table>
<thead>
<tr>
<th>Network</th>
<th>Evaluation</th>
<th>Accuracy</th>
<th>F1</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gertych et al.(^{68})</td>
<td>Cross-validation</td>
<td>-</td>
<td>-</td>
<td>0.595 ± 0.15</td>
</tr>
<tr>
<td>Li et al.(^{72})</td>
<td>Cross-validation</td>
<td>-</td>
<td>-</td>
<td>0.737</td>
</tr>
<tr>
<td>Our method</td>
<td>Hold-out validation</td>
<td>0.866 ± 0.07</td>
<td>0.835 ± 0.13</td>
<td>0.735 ± 0.16</td>
</tr>
</tbody>
</table>

\(^{*}\) Li et al. reported separate scores for segmenting benign and cancerous epithelium. The score displayed here is the average of those two.
Automated Gleason grading of prostate biopsies

Authors: Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, Geert Litjens

Adapted from: Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study

Published in: The Lancet Oncology (Volume: 21, issue: 2, February 2020)

DOI URL: doi.org/10.1016/S1470-2045(19)30739-9
Abstract

The Gleason score is the strongest correlating predictor of recurrence for prostate cancer, but has substantial inter-observer variability, limiting its usefulness for individual patients. Specialized urological pathologists have greater concordance; however, such expertise is not widely available. Prostate cancer diagnostics could thus benefit from robust, reproducible Gleason grading. We aimed to investigate the potential of deep learning to perform automated Gleason grading of prostate biopsies.

In this retrospective study, we developed a deep-learning system to grade prostate biopsies following the Gleason grading standard. The system was developed using randomly selected biopsies, sampled by the biopsy Gleason score, from patients at the Radboud University Medical Center (pathology report dated between Jan 1, 2012, and Dec 31, 2017). A semi-automatic labeling technique was used to circumvent the need for manual annotations by pathologists, using pathologists’ reports as the reference standard during training. The system was developed to delineate individual glands, assign Gleason growth patterns, and determine the biopsy-level grade. For validation of the method, a consensus reference standard was set by three expert urological pathologists on an independent test set of 550 biopsies. Of these 550, 100 were used in an observer experiment, in which the system, 13 pathologists, and two pathologists in training were compared with respect to the reference standard. The system was also compared to an external test dataset of 886 cores, which contained 245 cores from a different center that were independently graded by two pathologists.

We collected 5759 biopsies from 1243 patients. The developed system achieved a high agreement with the reference standard (quadratic Cohen’s kappa 0·918, 95% CI 0·891–0·941) and scored highly at clinical decision thresholds: benign versus malignant (area under the curve 0·990, 95% CI 0·982–0·996), grade group of 2 or more (0·978, 0·966–0·988), and grade group of 3 or more (0·974, 0·962–0·984). In an observer experiment, the deep-learning system scored higher (kappa 0·854) than the panel (median kappa 0·819), outperforming 10 of 15 pathologist observers. On the external test dataset, the system obtained a high agreement with the reference standard set independently by two pathologists (quadratic Cohen’s kappa 0·723 and 0·707) and within inter-observer variability (kappa 0·71).

Our automated deep-learning system achieved a performance similar to pathologists for Gleason grading and could potentially contribute to prostate cancer diagnosis. The system could potentially assist pathologists by screening biopsies, providing second opinions on grade group, and presenting quantitative measurements of volume percentages.
3.1 Introduction

With 1.2 million new prostate cancer cases each year worldwide, a high incidence-to-mortality ratio, and risk of overdiagnosis and overtreatment, an accurate assessment of patient prognosis is needed. The Gleason score, assigned by a pathologist after microscopic examination of cancer morphology, is the most powerful prognostic marker for patients with prostate cancer. However, substantial inter-observer and intra-observer variability in grading reduces its usefulness for individual patients. Specialized urological pathologists have greater concordance, but such expertise is not widely available. Prostate cancer diagnostics could thus benefit from robust, reproducible Gleason grading.

Treatment planning for prostate cancer is based mainly on the biopsy Gleason score. After the biopsy procedure, tissue specimens are formalin-fixed and paraffin-embedded, cut into thin sections, stained with hematoxylin and eosin, and examined under a microscope by a pathologist. The Gleason system stratifies the architectural patterns of prostate cancer into five types, from 1 (low risk) to 5 (high risk). The Gleason score in biopsies is the sum of the most common pattern and the highest secondary pattern (e.g., 3+5). Growth patterns 1 and 2 are not or rarely reported for biopsies.

In the latest revision of the Gleason grading system, five prognostically distinct grade groups were introduced; assigning scores 3+3 and lower to group 1, 3+4 to group 2, 4+3 to group 3, 3+5, 5+3, and 4+4 to group 4, and higher scores to group 5. Although clinically relevant, initial research shows that this transition has not reduced the observer variability of the grading system.

Artificial intelligence, particularly deep learning, has the potential to increase the quality of Gleason grading by improving consistency and offering expert-level grading independent of location. Deep learning has already been investigated and shown promising use in diagnostics in several medical fields, with examples in radiology, ophthalmology, dermatology, and pathology. For prostate cancer, previous studies have applied feature-engineering approaches to address Gleason grading. Eventually, the field transitioned to applications of deep learning for detecting cancer, and later Gleason grading of tissue microarrays, prostatectomies, and biopsies. Studies of biopsies have focused solely on Gleason 3 versus Gleason 4 in small datasets.

We aimed to produce a fully automated cancer detection and Gleason grading system for entire prostate biopsies, trained without the need for manual pixel-level annotations, focusing on the full range of Gleason grades, and evaluated on a large cohort of patients with an expert consensus reference standard, a separate observer study, and an external tissue microarray test dataset.
3.2 Methods

3.2.1 Study design and participants

For method development and validation, we retrospectively built several distinct datasets: the internal training, tuning, test, and observer datasets and the external training and test datasets (Figure 3.1).

From digital patient records of the Radboud University Medical Center, all pathologist reports dated between Jan 1, 2012, and Dec 31, 2017, for patients who underwent a prostate biopsy owing to a suspicion of prostate cancer were retrieved. The need for informed consent was waived by the local ethics review board (2016–2275). The reports were anonymized, and a text search was used to establish the highest mentioned Gleason score in each report. Patient reports were then randomly sampled using the train_test_split function of the scikit-learn Python package (version 0.20.2), stratifying by the Gleason score, resulting in an equal distribution of Gleason scores. Each pathology report was read, and for each patient, a single hematoxylin and eosin-stained glass slide containing the most aggressive or prevalent part of the tumor was selected for scanning. Additional reports mentioning only benign biopsies were selected. Patients who had neoadjuvant or adjuvant therapy were excluded. The resulting dataset is further referenced to as the internal dataset.

The selected glass slides were scanned using a 3DHistech Pannoramic Flash II 250 (3DHistech, Hungary) scanner at 20× magnification (pixel resolution 0.24 μm). Each scan contained one to six unique biopsies, commonly with two sections per biopsy. After scanning, trained non-experts assessed all slides and coarsely outlined each biopsy, assigning each with a Gleason score or labeling negative on the basis of the pathology report. A fixed number of slides were randomly assigned into datasets for testing or tuning, and the remainder were assigned to the training dataset. Randomization was stratified by patient and highest Gleason grade.

From the internal test dataset, a subset of 100 biopsies was selected to be presented to a group of pathologists in an observer experiment, further referenced to as the observer dataset. The size of the observer dataset was decided in consultation with experts (HvB, RV, and CHvdK). One of the expert pathologists (CHvdK) selected 20 benign cases manually, controlling for a broad range of tissue patterns, including inflammation and (partial) atrophy. The remaining 80 biopsies were randomly selected, stratified for Gleason grade group on the basis of the reported values of the same pathologist.

The 100 biopsies were made available through an online viewer, PMA.view (Pathomation, Berchem, Belgium), and distributed to an external panel. Panel members
3.2 Methods

Patient record query

Randomly sampled (Stratified for Gleason grade)

Scanning & annotating

Training slides (N=933)

Tuning slides (N=100)

Test slides (N=210)

Biopsies distributed to experts (N=550)

Biopsies graded by experts (N=535)

Biopsies graded by external pathologists (N=100)

Excluded by experts\(^5\) (N=15)

Annotated biopsies (N=4712)

Annotated biopsies (N=497)

Annotated slides (N=1243)

Scanned slides (N=1410)

Glass slides (N=1470)

Excluded (N=60):
- Glass slide not retrievable (N=11)
- Duplicate slide\(^1\) (N=49)
- Non-prostate cancer (N=2)
- Inconclusive report\(^2\) (N=96)
- Scan invalid / empty\(^3\) (N=49)
- Adjuvant therapy\(^4\) (N=12)
- Scanner missed biopsies (N=8)

Biopsies distributed to experts (N=550)

Excluded (N=167):
- Non-prostate cancer (N=2)
- Inconclusive report\(^2\) (N=96)
- Scan invalid / empty\(^3\) (N=49)
- Adjuvant therapy\(^4\) (N=12)
- Scanner missed biopsies (N=8)

Figure 3.1: Study profile

1 For some patients, glass slides were accidentally scanned twice. Duplicates were discarded, resulting in one included glass slide per patient.
2 For some cases, the trained non-experts were unable to match the scanned biopsies to the description from the pathologist report. Most commonly, if the report did not explicitly describe the individual biopsies on the glass slide.
3 Scans were excluded if the scanner failed to scan all or a majority of the tissue.
4 Some cases with adjuvant therapy were missed by the automated text-search.
5 Cases were excluded if at least one of the experts determined that the biopsy could not be reliably graded. Reasons: Biopsy out of focus or unsharp (N=5), biopsy mechanically damaged (N=1), IHC needed (N=3), error in loading file (N=1), tumour area too small to grade (N=3), serial section needed (N=1), image quality too low (N=1).
were invited to participate in this study at the United States and Canadian Academy of Pathology 2019 annual meeting in Washington, DC, USA (March 16–21, 2019). Interested pathologists were asked to report their current affiliation, their experience with Gleason grading, and the number of cases they viewed annually, and were subsequently asked to invite colleagues in their network who had experience in Gleason grading. All pathologists who graded all 100 biopsies were included. All panel members had experience with Gleason grading, but with a varying amount of experience. No time restriction was given, although we asked that they complete the grading within 6 weeks.

We also evaluated the system on an external, independent, public dataset of tissue microarrays87 to assess the robustness of the system to data from a different center (Department of Pathology and Molecular Pathology, University Hospital Zurich, Switzerland).89 One tissue core of a representative tumor area per patient was taken from an online database, and every sample that was assigned to the test dataset of Arvaniti and colleagues87 was used for validation, a total of 245 cores. The complete dataset consisted of 886 tissue cores, each corresponding to a single patient. The cases were prepared and stained in an independent lab and scanned using a different scanner. We had no influence on the composition of and made no changes to the external dataset.

### 3.2.2 Test methods

The data acquisition of the internal dataset resulted in outlined biopsies with a single label per biopsy. More detailed annotations were required to train the deep-learning system to segment individual glands. We preprocessed the biopsies of the training and tuning set in four steps (Figure 3.2). First, tissue was automatically distinguished from background using a tissue segmentation network.75 Second, within tissue areas, a trained tumor detection system38 was applied to define a rough outline of the tumor. The outlined tumor regions still contained large areas of stroma, inflammation, or other non-epithelial tissue. Third, to refine the tumor masks, each biopsy was processed by an epithelial tissue detection system,90 after which tissue that was detected as non-epithelial tissue was removed from the tumor mask. Finally, detected tumor tissue was assigned a label on the basis of the Gleason score retrieved from the pathology report. A description of the individual systems is provided in the supplementary methods.

We first trained a deep-learning system only on biopsies with a pure Gleason score (3+3, 4+4, or 5+5).91 After training, this initial system was applied to the internal training dataset to set the reference standard. By use of the pathologist reports,
1. Semi-automatic data labeling

a. 5759 prostate biopsies are used to develop the system. The training set is labeled semi-automatically.

b. First, a rough tumor outline is generated by a tumor detection system.

c. Non-epithelial tissue is then removed by an epithelium segmentation system.

d. The Gleason pattern from the pathologist’s report is assigned to the detected tumor area.

e. A system is trained on pure biopsies (3+3, 4+4, 5+5). After training, the system can segment patterns individually.

2. Refinement & training

f. The full training set is labeled using the network trained on pure biopsies. Reports are used to further refine the labels.

g. Using the new labels the final system is trained.

3. Grade group prediction

h. To evaluate, the final system is applied to the test set. Each gland is labeled with Gleason 3, 4, 5 or benign. The normalized percentages are used to compute the Gleason grade group.

Figure 3.2: Overview of the development of the deep learning system.
Automated Gleason grading of prostate biopsies

the output was automatically refined by removing clearly incorrect label assignments, such as cancerous glands in benign biopsies. Any tissue originating from benign biopsies detected as malignant was relabelled as hard negative (i.e., a sample of benign tissue that was difficult for the system to correctly classify) to be oversampled during training. A connected components algorithm, based on the \texttt{ndimage.label} function from the Python SciPy package (version 1.2.1), was applied to ensure that each gland was assigned to a single class.

The patients included in the internal test dataset were independent of the patients in the internal training and tuning datasets. To create a strong reference standard, we asked three pathologists with a subspecialty in urological pathology (CHvdK, HvB, and RV) to grade the biopsies individually through the online viewer, PMA.view, following the International Society of Urological Pathology 2014 guidelines. Clinical information of the patients was not available for the experts.

The reference standard for the internal test dataset was determined in three rounds. In the first round, each pathologist reviewed the biopsies individually. For positive biopsies, each pathologist was asked to report: primary, secondary, and tertiary Gleason grade (if present), total tumor volume, tumor volumes for the growth patterns, and the Gleason grade group. In the second round, each biopsy without consensus was regraded by the pathologist whose score differed from the other two. Additional to the pathologist’s initial examination, the Gleason scores of the other pathologists were appended anonymously. Biopsies without consensus after round two were discussed in a consensus meeting.

Our deep-learning system consisted of an extended U-Net\textsuperscript{52} that was trained on patches extracted from the internal training dataset. After the aforementioned semi-automatic labeling process, the system was trained on the complete training dataset, including biopsies with mixed Gleason growth patterns. The tuning dataset was used to monitor performance during training and to prevent overfitting. Training of the system was halted if no performance increase was measured on the tuning dataset.

The label quality of the internal training dataset was determined by labeling the test cases using the same automated method. We compared the retrieved Gleason scores of the test dataset with the final consensus score of the experts. The kappa values of this comparison acted as a measure of label quality.

After training, the deep-learning system was applied to all biopsies from the test dataset and compared with the reference standard. Test positivity cutoffs were determined before the analysis of the test dataset using the tuning dataset. The system determined the grade group of a biopsy in two steps. In the first step, the whole biopsy was segmented by assigning Gleason growth patterns to tumorous glands, and benign glands are classified as benign. From this segmentation, a normalized ratio
of epithelial tissue could be calculated as percentages of benign, grade 3, grade 4, or grade 5. Based on the tuning dataset, we classified a biopsy as malignant if at least 10% of the epithelial tissue was predicted as cancer by the system. In the second step, the grade group was determined on the basis of the normalized volume percentages of each growth pattern.

To apply the system to the external dataset and to account for stain and scanner variations, we applied an unsupervised normalisation algorithm based on CycleGANs. After normalization of the external test images, our deep-learning system, without any modification, was applied to the normalized test images. The reference standard for the external test dataset was based on the Gleason score. To account for this difference in test metrics, we determined test positive cutoffs on the external training data. For the external test dataset, no consensus score was available for the two pathologists who graded all cases; instead, we evaluated our method using both pathologists in turn as the reference standard.

### 3.2.3 Statistical analysis

After consultation with experts (HvB, RV, and CHvdK), 550 cases for the test dataset was established as a good balance between time investment and case diversity.

We defined the main metric as the agreement with the consensus reference standard, measured using quadratic Cohen’s kappa. To compare the performance of the system with the external panel of pathologists, we did multiple permutation tests. The test statistic was defined as the difference between the kappa of the deep-learning system and the median kappa of the pathologists. The analysis of the receiver operating characteristic curves was done using the difference in F1 score as test metric. Statistical analysis was done using Python 3.6 with the NumPy (1.16.3), pandas (0.25.1), scipy (1.2.1), scikit-learn (0.20.2) and matplotlib (2.2.4) packages. Further details are provided in the supplementary methods.

### 3.2.4 Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.
3.3 Results

3.3.1 Dataset & reference standard

After screening of 1329 slides from patients with malignant disease and 141 slides from patients with benign disease, 1410 slides were scanned. After exclusion of 167 slides, the internal dataset consisted of 1243 patients and 5759 biopsies. The training dataset consisted of 933 (75%) of 1243 slides (4712 biopsies), the tuning dataset of 100 (8%) of 1243 slides (497 biopsies), and the test dataset of 210 (17%) of 1243 slides (550 biopsies; Figure 3.1). The observer dataset was sampled from the test dataset and consisted of 100 biopsies from 78 patients.

After the first round of grading of biopsies in the internal test dataset, 333 (61%) of 550 cases had complete consensus. The three experts’ inter-rater agreement was high (quadratic Cohen’s kappa of 0.925). The majority vote was taken for some slides: 11 (2%) of 550 cases with an agreement on grade group, but a difference in Gleason pattern order (eg, 5+4 versus 4+5); 11 (2%) cases with an equal grade group, but a disagreement on Gleason score; and 110 (21%) cases for which two pathologists agreed and the third had a maximum deviation of one grade group. Cases with a disagreement on malignancy were always flagged for a second read. Immunohistochemistry was used in seven (1%) of 550 cases to determine the benign or malignant label as it was present in the original report. 15 (3%) of the 550 cases were excluded by the experts because they could not be reliably graded. In the second round, 63 (11%) of 550 cases were regraded. 27 (5%) of 550 cases of biopsies without consensus after round two were discussed in a consensus meeting. Grade group distribution and confusion matrices are presented in Supplementary Figure 3.3.

To get an estimate of label noise, the reference standard was compared with labels generated by the semi-automatic method; the accuracy of the retrieved labels versus the reference was 0.675 (kappa 0.819) for Gleason score and 0.720 (kappa 0.853) for grade group.

3.3.2 Performance on the internal test set

On the 535 biopsies of the internal test dataset, our deep-learning system achieved an agreement of 0.918 (quadratic Cohen’s kappa, 95% CI 0.891–0.941) with the consensus grade group. Most errors by the deep-learning system are made in distinguishing between grade group 2 and 3, and grade group 4 and 5 (Figure 3.3).

In the internal test dataset, the deep-learning system missed 13 malignant cases, of which all but one were determined as grade group 1 by the experts (Figure 3.3a). In
12 of these cases, the system detected a tumor, but the predicted volume was below our threshold for malignancy.

Receiver operating characteristic curve analysis on three clinically relevant cutoffs showed the ability of the system to group cases in risk categories with high accuracy.
Automated Gleason grading of prostate biopsies

Figure 3.4: Bootstrapped ROC analysis on tumor versus benign. The left column shows results on the internal test set (N=535). The right column shows the results on the observer set (N=100); the values for each panel member have been added to the graphs.

(Figure 3.4a, 3.5a and 3.6a; Table 2). The decision threshold of the system can be tuned to correctly predict 99% of biopsies containing tumor with a specificity of 82%.

3.3.3 Comparison to a panel of pathologists

For the observer study, 13 pathologists and two pathologists in training from 14 independent labs and ten countries individually graded all 100 biopsies following the International Society of Urological Pathology 2014 guidelines. This external panel showed a median inter-rater agreement of 0·819 (quadratic kappa, 95% CI 0·726–0·869) on Gleason grade group with the consensus (Figure 3.7). The system achieved a kappa value of 0·854 (quadratic kappa, 0·777–0·914) on the cases of the observer dataset, scoring higher than the median value of the panel and outperforming 10 of the 15 panel members (Figure 3.7). The performance of the deep-learning system was better than that of pathologists with less than 15 years of experience (two-sided permutation test, p=0·036) and scores not significantly different than pathologists with more than 15 years of experience (two-sided permutation test, p=0·96). To exclude bias towards the experts in our results, we computed the inter-rater agreement between all panel members independently of the reference standard. We then computed the agreement of the system with all members of the panel (Figure 3.1).
3.3 Results

![ROC analysis](image1)

**Figure 3.5:** Bootstrapped ROC analysis on benign and grade group 1 versus grade group ≥ 2. The left column shows results on the internal test set (N=535). The right column shows the results on the observer set (N=100); the values for each panel member have been added to the graphs.

![ROC analysis](image2)

**Figure 3.6:** Bootstrapped ROC analysis on low (benign and grade group 1-2) versus high grade cancer (grade group 3-5). The left column shows results on the internal test set (N=535). The right column shows the results on the observer set (N=100); the values for each panel member have been added to the graphs.
Sorted by the median kappa value, the deep-learning system had the third-highest inter-rater agreement score.

The deep-learning system scores better than three pathologists of the panel but lower than most on accuracy (Supplementary Figure 3.2). The lower accuracy is mostly caused by one-off errors between grade groups 2 versus 3 and 4 versus 5 (Figure 3.3). A two-sided permutation test on the difference between system accuracy and the median of the panel showed no significant difference (p=0·15). See Figure 3.8 for example cases.

After receiver operating characteristic analysis of the observer dataset, a two-sided permutation test on the median F1 score showed no difference between the deep-learning system and the panel for both malignant versus benign (p=0·70), grade group 2 as a cutoff (p=0·84), and grade group 3 as a cutoff (p=0·65; Figure 3.4b, 3.5b and 3.6b; Table 2).

### 3.3.4 Performance on the external test set

The external test dataset contained 245 cores that were independently graded by two pathologists (inter-rater agreement quadratic Cohen’s kappa 0·71). Concerning the two pathologists, the system obtained a 0·723 and 0·707 quadratic kappa on Gleason score. A receiver operating characteristic analysis for relevant decision thresholds was also done for the external test set (Figure 3.9). Overall, the deep learning system performed comparably to the two pathologists who set the reference standard.
Figure 3.8: Examples from the observer set with the consensus grade group, the predictions by the system, and the grade groups from the panel. In the zoomed regions the Gleason pattern prediction of the system is shown as an overlay on the tissue.
Automated Gleason grading of prostate biopsies

Figure 3.9: Bootstrapped ROC analysis on the external test set using three clinical relevant cutoffs: 1) tumor versus benign (top row); 2) benign and grade group 1 versus grade group $\geq 2$ (middle row); and 3) low (benign and grade group 1-2) versus high grade (grade group 3-5) cancer (bottom row). The predictions of the deep learning system are compared to the two pathologists that set the reference standard.
3.4 Discussion

We have developed a fully automated method to grade prostate biopsies and have shown that this method can achieve a performance similar to pathologists on both the internal and external test datasets. The performance of the deep-learning system could only reliably be assessed by use of an expert reference standard. We asked three expert urological pathologists to grade the complete test dataset, which resulted in a minimum of three independent reads for every case. The deep-learning system achieved a high agreement (quadratic kappa of 0.918) with the reference standard. We also compared the system with a panel of independent pathologists and pathologists in training. In this observer dataset, the deep-learning system outperformed ten of 15 panel members. On the external test dataset, the system showed it could generalize to external and unseen data. The system scored comparably to the results attained by Arvaniti and colleagues (quadratic kappa, 0.723 and 0.707 vs 0.71 and 0.75) and within inter-observer variability of the pathologists who set the reference standard (kappa 0.71), although our system was not trained on data from that set.

The training data was labelled in a semi-supervised way, saving resources that would otherwise have been spent in manually labeling slides. Moreover, it is often practically unfeasible to precisely annotate the vast amounts of data required for deep learning, even though unannotated or sparsely annotated data is often readily available in pathology archives. One limitation of this method is that it can introduce label noise in the training dataset. However, the ability of deep-learning systems to handle substantial amounts of label noise is well known.

The black-box characteristic of deep learning is often mentioned as a drawback of such systems, especially for medical decision making. We addressed this drawback by having our system show predictions at multiple abstraction levels, instead of using an additional learned model on top of the deep-learning system. The precise gland-level segmentations of the developed system made it possible to use a simple ruleset on grade volume percentages to obtain the biopsy-level grade, similar to the one described in the Gleason grading system. This approach allows pathologists to assess whether the epithelium was correctly classified, whether the system missed certain glands, and the grades assigned to individual or groups of glands. As such, our system provides a higher level of interpretability compared with competing approaches.

Given the high prevalence of prostate cancer, reducing workload for pathologists is of clinical value. In the test dataset, our deep-learning system achieved an AUC of 0.990 on determining the malignancy of a biopsy, on the observer set an AUC of 0.984. Furthermore, the system can be tuned to achieve a sensitivity of 99%. As such,
Automated Gleason grading of prostate biopsies

Figure 3.10: Failure cases from the observer set. For each case, the grade group of the reference standard, the prediction by the deep learning system, and the distribution of grade groups from the panel are shown. In the zoomed regions the Gleason pattern prediction of the system is shown as an overlay on the tissue.

Our system could be implemented as a prescreening triage tool within pathology labs, giving priority to high-grade biopsies and filtering out low-risk benign biopsies. More work is needed to increase the discrimination power of the system between Gleason grade groups 2 and 3 and groups 4 and 5. The boundaries between group 2 and 3 are defined by the relative volume percentages of the Gleason growth patterns, which makes an accurate estimate of those volumes essential for correct classification. Group 4 versus 5 is complicated by a wide range of Gleason scores that fall under these two groups (3+5, 5+3, 4+4, 4+5, 5+4, and 5+5). Discrepancies can also be related to the way the system and pathologists differ in estimating the relative volume of growth patterns. The system counts the exact area of the individual glands, whereas a pathologist assesses the volume more qualitatively. Some failure cases of the system are shown in Figure 3.10.

Our results extend previous work on prostate cancer detection\textsuperscript{38,86} and automated Gleason grading\textsuperscript{27,87,88}. We extend on these works by focusing on automated Gleason grading for prostate biopsies, the strongest histological correlating predictor of recurrence for patients with prostate cancer, of which the grading system differs from prostatectomies. Furthermore, by including both benign biopsies and biopsies from
the full spectrum of Gleason grades, we created a system that is usable as a prescreening tool and as a second reader.

Grading of the biopsies, both by the experts and the external panel, was done through digital viewing of the slides. For the external panel, not all members had previous experience with digital viewing or with the digital viewer used in this study. Owing to this inexperience, we cannot know whether it affected their grading. Nonetheless, research has shown that digital viewing is non-inferior to microscopy.\textsuperscript{93}

Before our system can be used in clinical practice, some limitations must be addressed. First, the data that were used to develop the deep-learning system originated from a single center. Although the performance on the external test dataset is within the range of inter-observer variability, including data from multiple centers, with different staining protocols and whole slide scanners, could further increase the robustness of the system. Second, we focused on the grading of acinar adenocarcinoma in prostate biopsies, although other tumor types and foreign tissue can be present in prostate biopsies (eg, colon glands, which should be identified and excluded for grading). Additionally, other prognostic information could be present in the biopsies that we did not extract (eg, the detection of intraductal carcinoma).\textsuperscript{94} Finally, in this study, each biopsy is treated independently, both by the pathologists and by the deep-learning system. In clinical practice, multiple biopsies are sampled from different regions of the prostate. An update to the deep-learning system could take multiple biopsies into account and give a grade group prediction at the patient level.

The developed system will be made available for scientific and non-commercial use, through the Radboudumc Computational Pathology Group website. Furthermore, through a grand challenge on Gleason grading, we will publish a part of our data for others to use in developing new methods for automated Gleason grading.

Our automated deep-learning system achieved a performance similar to pathologists in terms of Gleason grading. With further evaluation, the system could assist pathologists by screening biopsies, providing second opinions on grade group, and presenting quantitative measurements of volume percentages.

### 3.5 Acknowledgements

This study was funded by a grant from the Dutch Cancer Society (KWF), grant number KUN 2015-7970.

We would like to thank the following pathologists and pathologists in training for participating in our study as part of the panel:
We also thank Jeffrey Hoven for assisting with the data collection and scanning, and Milly van de Warenburg, Nikki Wissink, and Frederike Haverkamp for their help making the manual annotations.
3.6 Supplementary figures

3.6.1 Grade group agreement (quadratic kappa) between panel and deep learning system

**Supplementary Figure 3.1**: Inter-rater agreement between panel members. For each pathologist, the inter-rater agreement with each other pathologist from the panel was calculated. Additionally displayed is the agreement of the deep learning system with the pathologists from the panel. The pathologists and deep learning system are ordered based on their respective median agreement values. The two horizontal lines display the median agreement of all pathologists (in blue) and median agreement of the system (in green).
3.6.2 Accuracy deep learning system accuracy versus panel

Supplementary Figure 3.2: Inter-rater agreement (non-weighted accuracy measure) between external pathologists. For each pathologist the agreement with each other pathologist was calculated. Additionally displayed is the agreement of the deep learning system with all pathologists.
3.6.3 Confusion matrices experts with consensus on test set

Supplementary Figure 3.3: Confusion matrices for the three expert pathologists on the test set. Absolute numbers in a-c, normalized frequencies in d-f.
3.7 **Supplementary methods**

3.7.1 **Tumor detection system**

A previously developed tumor detection system was applied to outline tumor areas in our training set. To train this tumor detection system, a pathologist in training, supervised by an experienced uropathologist, outlined tumor regions in 100 prostate biopsies. Most of these biopsies were low grade: 52 benign, 11 grade group 1, 23 grade group 2, seven grade group 3, five grade group 4, and two biopsies grade group 5. The biopsies used for the development of the tumor detection system were independent of the biopsies used in the current study.

Patches were extracted from the 100 biopsies and used to train the system (pixel resolution of 1.92µm). The tumor detection system achieved an AUC of 0.99 in discriminating benign from malignant biopsies on a separate test set of 75 biopsies.

After training, the system was applied as a fully convolutional network to all biopsies of the current study. This procedure resulted in a rough outline of tumor regions.

3.7.2 **Epithelium segmentation system**

A previously developed epithelium segmentation system was used to refine the tumor outlines generated by the tumor detection system. The epithelium segmentation system was developed using 102 prostatectomy tissue sections. The tissue sections were stained with H&E and subsequently restained with P63 and CK8/18 immunohistochemistry (IHC) markers to highlight epithelial structures. Each H&E and IHC pair were subsequently co-registered.

An initial deep learning system was trained on a subset of the IHC slides that were preprocessed with color deconvolution. This trained system was then applied to all IHC slides in the training set, forming a reference standard for the final system. This automated labeling method made sure that even poorly differentiated Gleason 5 areas were precisely annotated.

The final epithelium segmentation system was trained on the H&E slides using the automatically generated reference standard. A five-level-deep U-Net was used as the network architecture with patches extracted at a pixel resolution of 0.98µm. The system achieved a high segmentation performance (F1 score of 0.893) and was able to segment both intact glands and individual malignant epithelial cells.
3.7.3 Overview of deep learning system

Our deep learning system consisted of a U-Net\textsuperscript{52} that was trained on randomly sampled patches extracted from the training set. After the automatic labeling process, the system could be trained on all biopsies, including those with mixed Gleason growth patterns. Additional patches were sampled from the hard-negative areas to improve the system’s ability to distinguish tumor from benign tissue.

We followed the original U-Net architecture but experimented with different configurations. Given the importance of morphological features in Gleason grading, we focused on multiple depths of the U-Net, combined with different pixel spacings for the training data. Experimentation showed the best performance on the tuning set using a depth of six levels, sampled patches with a size of $1024 \times 1024$, and a pixel resolution of $0.96 \mu m$. We added additional skip connections within each layer block and used up-sampling operations in the expansion path. Adam optimization was used with $\beta_1$ and $\beta_2$ set to 0.99, a learning rate of 0.0005 and a batch size of 8. The learning rate was halved after every 25 consecutive epochs without improvement on the tuning set. We stopped training after 75 epochs without improvement on the tuning set. Adding additional training examples from hard negative areas increased the performance of the network, especially in distinguishing between benign, inflammatory, and tumorous tissue.

The network was developed using Keras\textsuperscript{48} and TensorFlow.\textsuperscript{47} Data augmentation was used to increase the robustness of the network. The following augmentation procedures were used: flipping, rotating, scaling, color alterations (hue, saturation, brightness, and contrast), alterations in the H&E color space, additive noise, and Gaussian blurring.

3.7.4 Determining the Gleason grade group for a new specimen

Our deep learning system determines the Gleason grade group for a biopsy in two steps. First, our trained U-Net is applied to the scanned tissue of the biopsy. This procedure results in a label for each pixel of the image: background, stroma, benign epithelium, Gleason 3, Gleason 4, or Gleason 5. The frequency of each label can then be counted. Biopsies can differ vastly in size and in the amount of epithelial tissue that is present. To account for this difference, the values for the three Gleason growth patterns are normalized based on the sum of benign and malignant epithelial tissue. By normalizing, we obtain a volume estimate of the tumor that is independent of the size of the biopsy.

The volume percentages were used to assign a Gleason score and Gleason Grade group based on the guidelines for biopsies in clinical practice.\textsuperscript{19} First, we determine
whether a biopsy is malignant or benign. Based on the tuning set, we classify a biopsy as malignant if at least 10% of the epithelial tissue is predicted as cancer by the system. For malignant biopsies, we then determine the Gleason score. The growth pattern that has the largest volume is taken as the primary component. If there are other growth patterns present, with a volume of at least 7%, the most aggressive component is used as the secondary pattern. The 7 and 10% cut-offs were determined automatically based on the tuning set. Note that this procedure differs between prostatectomies and biopsies. For prostatectomies, the secondary pattern is always the second-largest growth pattern, regardless of aggressiveness. The predicted Gleason score is used to determine the grade group. A Gleason score 3+3 is mapped to group 1; Gleason score 3+4 is mapped to group 2; Gleason score 4+3 is mapped to group 3; Gleason scores 3+5, 4+4 and 5+3 are mapped to group 4; and higher scores are mapped to Gleason 5.

3.7.5 CycleGAN for style transformation and application to external data

We used a cycle-consistent generative adversarial network (CycleGAN) system to facilitate stain transformation on the external dataset of tissue microarrays. In a CycleGAN setup, two separate networks are trained to perform a transformation from one stain to the other, while retaining the structural information of the tissue. We used a previously developed CycleGAN setup that was developed for the transformation of histopathological tissue. The CycleGAN was adapted to learn the residual change instead of reconstructing the full image. To train the CycleGAN system, we sampled patches from the internal training set and used the full external dataset. Before inference with the Gleason deep learning system, we applied the CycleGAN network on the whole external dataset. The CycleGAN network was implemented in TensorFlow.

3.7.6 Determining Gleason score for the external dataset

The original paper of the external dataset reports the quadratic kappa on Gleason score as the primary metric. The Gleason scores were determined using the standard for grading prostatectomies: the sum of the most and second most common growth patterns. We applied our algorithm to the test set of this paper and computed the quadratic kappa on Gleason score to allow a one-to-one comparison of our algorithm to the algorithm developed by Arvaniti et al. To account for the difference in the grading systems between prostatectomies and biopsies, we adjusted the decision thresh-
olds of the deep learning system using the training data set of 641 cores, resulting in a tumor threshold of 3% and a secondary pattern threshold of 1.5%.

### 3.7.7 Statistical analysis

To compare the performance of the deep with the external panel of pathologists, we performed multiple permutation tests. The permutation test was implemented as follows: For each case in the observer set, we obtained a list of predictions, one by the deep learning system and the remainder by the pathologists. In each iteration of the permutation test, for each case, we swapped the grade group prediction of the system with a random prediction from the list of predictions. After swapping the predictions, the test statistic was computed. Repeating this procedure 10,000 times resulted in a null distribution of the test statistic. The original test statistic was then compared to the null distribution, resulting in a two-tailed p-value.

We defined the main metric as the agreement with the consensus reference standard, measured using quadratic Cohen’s kappa. For the analysis on agreement with the reference, we defined the test statistic as the difference between the kappa of the deep learning system and the medium kappa of the pathologists. The panel of pathologists was split into two groups, those with less than 15 years experience and those with more, and a permutation test was performed for both groups. The analysis of the ROC curves was done using the difference in F1-score as the metric. The decision threshold for the system was based on the point that maximized the AUC. The comparison of grade group accuracy was made using the difference in the accuracy of the system and the median accuracy of the pathologists with respect to the reference standard.
The PANDA Challenge

Authors: Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Henrik Grönberg, Hemamali Samaratunga, Brett Delahunt, Toyonori Tsuzuki, Tomi Häkkinen, Lars Egevad, Maggie Demkin, Sohier Dane, Lily Peng, Craig H. Mermel, Pekka Ruusuvuori, Geert Litjens, Martin Eklund, and the PANDA Challenge consortium

Original title: Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer in Biopsies: The PANDA Challenge

Accepted for publication after revision in: Nature Medicine
Abstract

Artificial intelligence (AI) has shown promise for diagnosing prostate cancer in biopsies. However, results have been limited to individual studies, lacking validation in multinational settings. Competitions have shown to be accelerators for medical imaging innovations, but their impact is equally hindered by lack of reproducibility and independent validation. With this in mind, we organized the PANDA challenge – the largest histopathology competition to date, joined by 1,290 developers – to catalyze development of fully reproducible AI algorithms for Gleason grading using 10,616 digitized prostate biopsies. We validated that a diverse set of submitted algorithms reach pathologist-level performance on independent cross-continental cohorts, fully blinded to the algorithm developers. The results warrant evaluating AI in prospective clinical trials.
4.1 Introduction

Gleason grading of biopsies yields important prognostic information for prostate cancer patients and is a key element for treatment planning. Pathologists characterize tumors into different Gleason growth patterns based on the histologic architecture of the tumor tissue. Based on the distribution of Gleason patterns, biopsy specimens are categorized into one of five International Society of Urological Pathology (ISUP) grade groups. This assessment is inherently subjective with considerable inter- and intra-pathologist variability, leading to both undergrading and overgrading of prostate cancer.

Individual artificial intelligence (AI) algorithms have shown promise in grading prostate cancer in biopsies and assisting pathologists in the microscopic reviews. However, AI algorithms are susceptible to various biases in their development and validation. This can result in algorithms that perform poorly outside the cohorts used for their development. Moreover, shortcomings in validating the algorithms’ performance on additional cohorts may lead to such deficiencies in generalization to go unnoticed. Algorithms are also often developed and validated in a siloed manner: the same researchers who develop the algorithms also validate them, introducing a risk of positive bias. This includes establishing the validation cohorts and selecting the pathologists providing the reference standard. There has yet to be an independent evaluation of algorithms for prostate cancer diagnosis and grading to assess whether they generalize across different patient populations, pathology labs, digital pathology scanner providers, and across reference standards derived from intercontinental panels of uropathologists. This represents a key barrier to implementation of algorithms in clinical practice.

AI competitions have been an effective approach to crowd-source the development of performant algorithms. Despite their effectiveness in facilitating innovation, competitions still tend to suffer from a set of limitations. Validation of the resulting algorithms has typically not been performed independently of the algorithm developers. In a competitive setup, the incentive for conscious or subconscious introduction of positive bias by the developers is arguably further increased, and lack of independent validation also means that the technical reproducibility of the proposed solutions is not verified. Moreover, competitions have typically not been followed up by validation of the algorithms on additional international cohorts, casting doubts on whether the resulting solutions possess the generalization capability to truly answer the underlying clinical problem, as opposed to being fine-tuned for a particular competition design and dataset.
Through this study, we aimed to advance the methodology for the design and evaluation of medical imaging AI innovations to develop and rigorously validate the next generation of algorithms for prostate cancer diagnostics. We organized a global AI competition, the Prostate CANcer graDe Assessment (PANDA) Challenge, by compiling and publicly releasing a European (EU) cohort for AI development, the largest publicly available dataset of prostate biopsies as of yet. Secondly, we fully reproduced top-performing algorithms’ and externally validated their generalization to independent United States (US) and EU cohorts and compared them to pathologists. The competition setup isolated the developers from the independent evaluation of the algorithms’ performance, minimizing the potential for information leakage and offering a true assessment of the diagnostic power of these techniques. Taken together, we show how the combination of AI and innovative study designs, together with pre-specified and rigorous validation across diverse cohorts can be utilized to solve challenging and important medical problems.

4.2 Results

4.2.1 Characteristics of the datasets

In total, 12,625 whole-slide images (WSI) of prostate biopsies were retrospectively collected from 6 different sites for algorithm development, tuning, and independent validation (Table 4.1, Supplementary Figures 4.1 to 4.4). Of these, 10,616 biopsies were available for model development (the development set), 393 for performance evaluation during the competition phase (the tuning set), 545 as the internal validation set in the post-competition phase, and 1,071 for external validation.

Cases for development, tuning and internal validation originated from Radboud University Medical Center, Nijmegen, The Netherlands and Karolinska Institutet, Stockholm, Sweden (Supplementary Figures 4.1 and 4.2 and Methods 4.7.1 and 4.7.2). The external validation data consisted of a US and an EU set (Supplementary Figures 4.3 and 4.4). The US set contained 741 cases and was obtained from two independent medical laboratories and a tertiary teaching hospital. The EU external validation set contained 330 cases and was obtained from the Karolinska University Hospital, Stockholm, Sweden. The histological preparation and scanning of the external validation samples were performed by different laboratories than those responsible for the development, tuning and internal validation data.
Figure 4.1: Overview of the PANDA challenge and study setup. The global competition attracted participants from 65 countries (top; size of the circle for each country illustrates the number of participants). The study was split into two phases. First, in the development phase (bottom left), teams competed in building the best performing Gleason grading algorithm, having full access to a development set for algorithm training, and limited access to a tuning set for estimating algorithm performance. In the validation phase (bottom right), a selection of algorithms was independently evaluated on internal and external datasets against reference grading obtained through consensus across expert uropathologist panels, and compared to groups of international and US general pathologists on subsets of the data.
<table>
<thead>
<tr>
<th>Source</th>
<th>EU Development set</th>
<th>EU Tuning Set</th>
<th>EU Internal Validation set</th>
<th>US External Validation set</th>
<th>US Internal Validation set</th>
<th>EU Development set</th>
<th>EU Tuning Set</th>
<th>EU Internal Validation set</th>
<th>US External Validation set</th>
<th>US Internal Validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radboud Univ. Medical Cent., Netherlands</td>
<td>5160 (9%)</td>
<td>330 (6%)</td>
<td>74 (2%)</td>
<td>2548 (42%)</td>
<td>237 (6%)</td>
<td>1028 (19%)</td>
<td>672 (13%)</td>
<td>2500 (45%)</td>
<td>3151 (55%)</td>
<td>330 (6%)</td>
</tr>
<tr>
<td>Karolinska Institutet, Sweden</td>
<td>546 (10%)</td>
<td>487 (10%)</td>
<td>146 (3%)</td>
<td>1262 (21%)</td>
<td>98 (2%)</td>
<td>672 (13%)</td>
<td>2500 (45%)</td>
<td>3151 (55%)</td>
<td>330 (6%)</td>
<td>2500 (45%)</td>
</tr>
<tr>
<td>Medical Lab., CA/UT, USA; Tertiary Teaching Hosp., CA, USA</td>
<td>5160 (9%)</td>
<td>330 (6%)</td>
<td>74 (2%)</td>
<td>2548 (42%)</td>
<td>237 (6%)</td>
<td>1028 (19%)</td>
<td>672 (13%)</td>
<td>2500 (45%)</td>
<td>3151 (55%)</td>
<td>330 (6%)</td>
</tr>
</tbody>
</table>

**Table 4.1: Data characteristics of the development set, tuning set, internal validation set, and the two external validation sets.** The development set was available to competition teams for algorithm development, and the tuning set for limited algorithm evaluation during the competition. All validation sets were fully independent and blinded to the algorithm developers. Additional details on the reference standard protocol in Supplementary Methods S2 and S3.
4.2 Results

4.2.2 Reference standards of the datasets

The reference standard for the Dutch part of the internal validation set was determined through consensus of 3 uropathologists (C.H.v.d.K., R.V., H.v.B.) from 2 institutions with 18 to 28 years of clinical experience after residency (mean, 22 years). For the Swedish subset, 4 uropathologists (L.E., B.H., H.S., T.T.) from 4 institutions, all with over 25 years of clinical experience after residency, set the reference standard. For the US external validation set, the reference standard was set through majority voting by 6 US or Canadian uropathologists (M.A., A.E., T.v.d.K., M.Z., R.A., P.H.) from 6 institutions with 18 to 34 years of clinical experience after residency (mean, 25 years). For this external dataset, immunohistochemistry was available to aid in tumor identification. The EU external validation set was reviewed by a single uropathologist (L.E.). For details on the uropathologist review protocol, see Supplementary Methods 4.7.2.

4.2.3 Overview of the competition

The study design of the PANDA challenge was pre-registered and consisted of a competition and a validation phase (Figure 4.1). The competition was open to participants from April 21st until July 23rd, 2020 and was hosted on the Kaggle platform (Supplementary Methods 4.7.5). During the competition phase, 1,010 teams consisting of 1,290 developers from 65 countries participated and submitted at least one algorithm. Throughout the competition, teams could request a benchmark of their algorithm on the tuning set, and were then simultaneously blindly validated on the internal validation set (Figures 4.2 and 4.3). All teams combined submitted 34,262 versions of their algorithms, resulting in a total of 32,137,756 predictions made by the algorithms.

The first team achieving an agreement with the uropathologists of above 0.90 (quadratically weighted Cohen’s kappa) on the internal validation set already occurred within the first 10 days of the competition (Figure 4.2a). In 33 days of the competition, the median performance of all teams exceeded a score of 0.85 (Figure 4.2b).

4.2.4 Overview of evaluated algorithms

After the competition, teams were invited to join the PANDA consortium. Of all eligible teams, 33 submitted a proposal to join the validation phase of the study. From these, the competition organizers selected 15 teams based on their algorithm’s performance on the internal validation set and method description (Supplementary Methods 4.7.6). Among the 10 highest-ranking teams in the competition, 8 submitted
Figure 4.2: Progression of algorithms’ performances throughout the competition. During the competition teams could submit their algorithm for evaluation on the tuning set, after which they received their score. At the same time, algorithms were evaluated on the internal validation set, without disclosing these results to the participating teams. The development of the top score obtained by any team (a) and the median score over all daily submissions (b) throughout the timeline of the competition shows the rapid improvement of the algorithms.
Figure 4.3: Final scores on the tuning (public leaderboard) and internal validation set (private leaderboard) for all participating teams at the end of the competition. A large fraction of teams reached high scores in the range of 0.80-0.90, and retained their performance on the internal validation set.

a proposal and were accepted to join the consortium. The further 7 teams in the consortium all ranked within the competition’s top 30.

All selected algorithms made use of deep learning-based methods\textsuperscript{43,81} Many of the solutions demonstrated the feasibility of end-to-end training using case-level information only\textsuperscript{109}, that is, using the ISUP grade group of a specimen as the target label for an entire WSI. Another algorithmic feature adopted by several top-performing teams was to apply automated label cleaning, where samples considered as erroneously graded by the pathologists were either excluded from training or relabeled. For a summary and details on the individual algorithms see Supplementary Methods 4.7.7.

4.2.5 Classification performance in the internal validation dataset

In the validation phase, all selected algorithms were fully reproduced on two separate computing platforms. The average agreement of the selected algorithms with the uropathologists was high with a quadratically weighted kappa of 0.931 (95% CI, 0.918-0.944, Figure 4.4a). Algorithms showed high sensitivity for tumor detection, with the representative algorithm (see Methods) achieving a sensitivity of 99.7\% (95% CI, 98.1-99.7, Figure 4.5) and a specificity of 92.9\% (95% CI, 91.9-96.7). The
classification performances of the individual algorithms are presented in Supplementary Figures 4.5 to 4.7.

4.2.6 Classification performance in the external validation datasets

The algorithms were independently evaluated on the two external validation sets. The agreements with the reference standards were high with a quadratic-weighted kappa of 0.862 (95% CI, 0.840-0.884) and 0.868 (95% CI, 0.835-0.900) for the US and EU external validation sets, respectively. The main algorithm error mode was overdiagnosing of benign cases as ISUP grade group 1 cancer (Supplementary Figures 4.8 and 4.9). The classification performances of the individual algorithms are presented in Supplementary Figures 4.5, 4.6, 4.8 and 4.9.

The representative algorithm identified cases with tumor in the external validation sets with sensitivities of 98.6% (95% CI, 97.6-99.3) and 97.7% (95% CI, 96.2-99.2). In comparison to the internal validation set, the algorithms misclassified more benign cases as malignant resulting in specificities of 75.2% (95% CI, 66.8-80.0) and 84.3% (95% CI, 70.5-87.9) for the representative algorithm.

4.2.7 Classification performance compared to pathologists

To compare algorithms’ performances to general pathologists, we obtained reviews from two panels of pathologists on subsets of the internal and US external validation sets. For the Dutch part of the internal validation set, 13 pathologists from 8 countries (7 from Europe and 6 outside of Europe) reviewed 70 cases. For the US external validation set, 20 US board-certified pathologists reviewed 237 cases. For details on the pathologist review protocol, see Supplementary Methods 4.7.3.

The algorithms scored significantly ($p < 0.001$) higher in agreement with the uropathologists (0.876, 95% CI, 0.797-0.927, Figure 4.4b) than the international general pathologists did (0.765, 95% CI, 0.645-0.852). The representative algorithm had higher sensitivity for tumor (98.2%, 95% CI, 97.4-100.0) than the representative pathologist (96.5%, 9% CI, 95.4-100.0) and higher specificity (100.0%, 95% CI, 90.6-100.0, versus, 92.3%, 95% CI, 77.8-97.8). On average, the algorithms missed 1.0% of cancers, whereas the pathologists missed 1.8%. Differences in grade assignments between the algorithms and pathologists are visualized in Figure 4.6.

On the subset of the US external validation set with pathologist reviews, the algorithms exhibited a similar level of agreement with the uropathologists as the US general pathologists did (0.828, 95% CI, 0.781-0.869 vs. 0.820, 95% CI, 0.760-0.865; $p = 0.53$). The representative algorithm had higher sensitivity for tumor (96.4%, 95% CI, 96.6-99.5) than the representative pathologist (91.9%, 95% CI, 89.3-95.5).
Figure 4.4: Algorithm agreement with reference standards and comparison to pathologists. Algorithms’ agreement (quadratically weighted kappa) with reference standards established by uropathologists is shown for the internal and external validation sets (left). On subsets of the internal and US external validation sets, agreement of general pathologists with the reference standards is additionally shown for comparison (right).
but lower specificity (75.0%, 95% CI, 61.2-82.7 versus 95.0%, 95% CI, 87.4-98.1). On average, the algorithms missed 1.9% of cancers, whereas the pathologists missed 7.3%. Differences in grade assignments between the algorithms and pathologists are visualized in Figures 4.7 and 4.8.

4.3 Discussion

AI has shown promise for diagnosis and grading of prostate cancer, but these results have been restricted to siloed studies with limited proof for generalization across diverse multinational cohorts, representing one of the central barriers to implementation of AI algorithms in clinical practice. The objective of this study was to overcome these critical issues. First, we aimed to facilitate community-driven development of AI algorithms for cancer detection and grading on prostate biopsies. Second, we sought to transcend isolated assessment of the diagnostic performance of individual AI solutions by focusing on reproducibility and fully blinded validation of a diverse group of algorithms on intercontinental and multinational cohorts.

The resulting PANDA challenge was the largest competition in pathology organized to date, both in terms of the number of participants and size of the datasets, and the first study to analyze a variety of AI algorithms for computational pathology at this scale. The datasets included variability in biopsy sampling procedure, specimen preparation process and whole slide scanning equipment, and had different sets of pathologists contributing to the reference standard. Our main finding was that AI algorithms obtained from a competition setup could successfully detect and grade tumors, reaching pathologist-level concordance with expert reference standards. Subsets of the internal and external validation datasets were also reviewed by groups of international and US pathologists. The algorithms had a concordance with the expert uropathologists that was similar to or higher than that of these groups of pathologists. In the external validation sets, the main algorithm error mode was overdiagnosing of benign cases as ISUP grade group 1. This is likely due to the data distribution shift between training data and external validation data in combination with the study design of independent validation, where the teams did not have any access to the validation sets, potentially leading to suboptimal selection of operating thresholds based only on the tuning set. We observed this in the US external validation set (Figure 4.5), where the algorithms appear to be shifted toward higher sensitivity but lower specificity compared to the general pathologists. A potential solution to address the natural data distribution shift is to calibrate the models’ predictions using sampled data from the target sites.
4.3 Discussion

Figure 4.5: Algorithm performance in detecting prostate tumors on validation sets. The sensitivity and specificity of the algorithms relative to reference standards established by uropathologists are shown for the internal and external validation sets (left). On subsets of the internal and US external validation sets, the sensitivity and specificity of general pathologists are also shown for comparison (right).
In the US external validation set, tumor identification was confirmed by immunohistochemistry, supporting the finding that the algorithms missed fewer cancers than the pathologists. This higher sensitivity shows promise for reducing pathologist workload by automated identification and exclusion of the majority of benign biopsies from review. Analysis of an ensemble constructed from the algorithms suggests that combining existing algorithms could improve specificity (Supplementary Table 4.3).

Further analysis of the grade assignments by the algorithms and general pathologists showed that the algorithms tended to assign higher grades than the pathologists (Figures 4.6 to 4.8). For example, in the US external validation set, algorithms overgraded a significant portion of ISUP grade group 3 cases as grade group 4. The general pathologists, in contrast, tended to under-grade cases, most notably in the high-grade cases. These differences suggest that general pathologists supported by AI could reach higher agreements with uropathologists, potentially alleviating some of the rater disagreement associated with Gleason grading.

We aimed to lower the entry barrier to medical AI development by providing access to a large, curated dataset, typically only attainable through large research consortia, and by organizing this competition to facilitate joint development with experience sharing among the teams. The results show that the publication of such datasets can lead to rapid development of high performing AI algorithms. Dissemination and fast iteration of new ideas resulted in the first team achieving pathologist-level performance in the first 10 days of the challenge (Figure 4.2a). These results show the important role data plays in the development of medical AI algorithms, given the short lead-time of top-performing solutions by various teams. At the same time, often raised criticisms of medical AI challenges are the lack of detailed reporting, and limited interpretation and reproducibility of results. Typically algorithms are only evaluated on internal competition data and by participants themselves, which introduces a risk of overfitting and reduced likelihood of reproducibility. We addressed these limitations in our challenge design by using pre-registration, blinded evaluation, full reproduction of algorithm results, independent validation of algorithms on external data and comparison with pathologists.

This study has limitations. First, algorithm validation focused on the assessment of individual biopsies whereas in clinical practice pathologists will examine multiple biopsies per patient. Second, this study focused on grading acinar adenocarcinoma of the prostate, and algorithm responses to other variants and subtypes of cancer, precancerous lesions, or non-prostatic tissue were not specifically assessed. Third, all the data were collected retrospectively across the institutions and the general pathologist reviews were conducted in a nonclinical setting, without additional clinical information available at the time of review. Fourth, despite the international nature of our
4.3 Discussion

evaluation (both in terms of pathologists’ practice and data sources), the countries involved were predominantly white, and demographic characteristics were not available in this study. Further investigation is required to validate the use of AI algorithms in more diverse settings. Lastly, this study did not evaluate the algorithm grading’s association directly with radical prostatectomy or clinical outcomes.

We found that a group of AI Gleason grading algorithms developed during a global competition generalized well to intercontinental and multinational cohorts with pathologist-level performance. On all external validation sets, the algorithms achieved high agreement with uropathologists and high sensitivity for malignant biopsies. The performance exhibited by this group of algorithms adds evidence of the maturity of AI for this task and warrants evaluation of AI for prostate cancer diagnosis and grading in prospective clinical trials. To stimulate further advancement of the field, the full development set of 10,616 biopsies is made publicly available.
Figure 4.6: ISUP grade group assignment by algorithms and pathologists. Algorithms are compared to international general pathologists on a subset of the internal validation set. Cases are ordered primarily by the reference ISUP grade group and secondarily by the average grade group of the algorithms and pathologists. Algorithms and pathologists are ordered by their agreement (quadratically weighted kappa) with the reference standard.
4.3 Discussion

Figure 4.7: ISUP grade group assignment by algorithms and pathologists. Algorithms are compared to international general pathologists on a subset of the US external validation set. Cases are ordered primarily by the reference ISUP grade group and secondarily by the average grade group of the algorithms and pathologists. Algorithms and pathologists are ordered by their agreement (quadratically weighted kappa) with the reference standard.
Figure 4.8: ISUP grade group assignment by algorithms and pathologists. Algorithms are compared to international general pathologists on a subset of the US external validation set. Cases are ordered primarily by the reference ISUP grade group and secondarily by the average grade group of the algorithms and pathologists. Algorithms and pathologists are ordered by their agreement (quadratically weighted kappa) with the reference standard. (Continuation of Figure 4.7.)
4.4 Methods

4.4.1 Study design

The study design of the PANDA challenge was pre-registered. We retrospectively obtained, and de-identified digitized prostate biopsies with associated diagnosis from pathology reports from Radboud University Medical Center, Nijmegen, The Netherlands and Karolinska Institutet, Stockholm, Sweden (Supplementary Figure S1 and Methods S1-S3). At the beginning of the competition, participating teams gained access to this EU development set of 10,616 biopsies from 2,113 patients for training of the AI algorithms (Table 4.1, Supplementary Methods 4.7.4). During the course of the competition, the teams could upload their algorithms to the Kaggle platform (Supplementary Methods 4.7.5) and receive performance estimates on a tuning set of 393 biopsies. Processing time was limited to 6 hours and the maximum GPU memory available was 16GB.

By the competition closing date, each team picked two algorithms of their choice for their final submission, and the higher-scoring of the two determined the team's final ranking. The final evaluation was performed on an internal validation dataset of 545 biopsies, collected from the same sites as the development and tuning sets and fully blinded to the participating teams. Moreover, to obtain an independent internal validation set, all samples from a given patient were used for either development or validation.

After the competition on the Kaggle platform ended, teams were invited to send in a proposal to join the validation phase of the study as members of the PANDA consortium. Joining the validation phase was fully voluntary and not a prerequisite for partaking in the competition. As a result, 15 teams were selected for further evaluation on two external validation datasets consisting of 741 and 330 biopsies, also fully blinded to the participating teams (Supplementary Methods 4.7.6 and 4.7.7). The first external validation set was obtained from two independent medical laboratories and a tertiary teaching hospital in the United States. The second external validation set was obtained from the Karolinska University Hospital, Stockholm, Sweden. All datasets consisted of both benign biopsies and biopsies with various ISUP grade groups. For details on the inclusion and exclusion criteria, see Supplementary Methods 4.7.1 and Supplementary Figures 4.1 to 4.4.

Whole slide images (WSI) of the biopsies were obtained using four different scanner models from three vendors: 3DHISTECH, Budapest, Hungary; Hamamatsu Photonics, Hamamatsu, Japan; and Leica Biosystems, Wetzlar, Germany (Supplementary Tables 4.1 and 4.2).
The study was approved by the institutional review board of Radboud University Medical Center (IRB 2016–2275), Stockholm regional ethics committee (permits 2012/572-31/1, 2012/438-31/3, and 2018/845-32), and Advarra (Columbia, MD; Pro00038251). Informed consent was waived due to the usage of de-identified prostate specimens in a retrospective setting.

### 4.4.2 Reproducing algorithms and application to validation sets

All teams selected for the PANDA consortium were asked to provide all data and code necessary for reproducing the exact version of their algorithm that resulted in the final competition submission. For each algorithm, we collected the main Jupyter notebook or python script for running the inference, the specific Kaggle Docker image (https://github.com/Kaggle/docker-python) used by the team during the competition, and any necessary associated files, including model weights and auxiliary code.

We replicated the computational setup of the competition platform and ran the algorithms on two different computational systems: Google Cloud and Puhti compute cluster (CSC - IT Center for Science, Espoo, Finland). On the Google Cloud platform, all algorithms were run using the original Docker images. On Puhti, the Docker images were automatically converted for use with Singularity. The algorithms and scripts provided by the teams were not modified except for minor adjustments required for successful run-time installation of dependencies on our computational systems. On Puhti, the algorithms had access to 8 CPU cores, 32 GB of memory, one Tesla V100 32GB GPU (Nvidia, Santa Clara, CA, USA) and 500 GB of SSD storage. On the Google Cloud platform, the algorithms had access to 8 CPUs, 30 GB memory, one Tesla V100 32GB GPU and 10000 GB of HDD storage.

Before applying the algorithms on the external validation sets, we first validated that the Kaggle computational environment had been correctly replicated and the algorithms' performance on our systems remained identical. To this end, we ran all algorithms on the tuning and internal validation sets on the two systems to reproduce the output generated during the competition on the Kaggle platform. By cross-checking the new results with the competition leaderboard we additionally assured that the algorithms supplied by the teams were not altered after the competition or tuned to perform better on the external validation sets. The verification runs we performed on the Puhti cluster were used as the basis for all results reported on the internal validation set.
Some algorithms were non-deterministic, e.g., because of test time augmentations with non-frozen random seeds. We ran each of these algorithms five times, and averaged the computed metrics.

After verification, we ran all algorithms on the external validation sets. For the US external validation set, we used the Google Cloud platform. For the EU external validation set, we used the Puhti cluster. This process was done independently of the teams and no prior information about the external datasets was supplied to the teams. The ISUP grade group predictions of the algorithms on the cases were saved and used as input for the analysis.

4.4.3 Statistical analysis

We defined the main metric as the agreement on ISUP grade group with the reference standard of each particular validation set, measured using quadratic Cohen’s kappa. To compare the performance of the algorithms with the general pathologists, we performed a two-sided permutation test per pathologist panel. The average agreement was calculated as the mean of the kappas across the algorithms and the pathologists, respectively. The test statistic was defined as the difference between the average algorithm agreement and the average pathologist agreement.

We calculated sensitivity and specificity on benign versus cancer-containing biopsies for all algorithms and individual general pathologists, based on the reference standard set by the uropathologists. Additionally, we determined the average number of biopsies with missed cancer per group. To further understand how a representative pathologist and algorithm performed, we selected the pathologist and the algorithm with the median balanced accuracy (the average of sensitivity and specificity), and reported the associated sensitivity and specificity of these two representatives. We reported the 95% confidence interval of the algorithms’ and pathologists’ performance metrics using bootstrapping, with both the algorithm or pathologist and case as the resampling unit.

Analysis was performed using Python (version 3.8) in combination with the following software packages: scipy (1.5.4), pandas (1.1.4), mlxtend (0.18.0), numpy (1.19.4), scikit-learn (0.23.2), matplotlib (3.3.2), jupyterlab (2.2.9) and notebook (6.1.5).

4.4.4 Data availability

The full development set, from here on named the PANDA Challenge dataset, of 10,616 digitized de-identified H&E stained prostate biopsies (383GB) will be made publicly available for further research. The data can be used under a Creative Com-
mons BY-SA-NC 4.0 license. The most up-to-date information regarding the dataset can be found on the challenge website at https://panda.grand-challenge.org/.

### 4.4.5 Code availability

Code that was used to generate the results of the various algorithms, and example code on how to load the images in the PANDA dataset is available at https://github.com/DIAGNijmegen/panda-challenge. Algorithms were built using open source deep learning frameworks, including Pytorch (https://pytorch.org/) and TensorFlow (https://www.tensorflow.org/). The Docker image that all the algorithms were based on is available online at https://github.com/Kaggle/docker-python.

### 4.5 Acknowledgements

Supported by the Dutch Cancer Society (KWF, grant number KUN 2015-7970), Netherlands Organization for Scientific Research (NWO) (project number 016.186.152), Google LLC, Verily Life Sciences, Swedish Research Council, Swedish Cancer Society, Swedish eScience Research Center, EIT Health, Karolinska Institutet, Åke Wiberg Foundation, Prostatacancerförbundet, Academy of Finland, Cancer Foundation Finland, ERAPerMed.

We thank the MICCAI board challenge working group, the MICCAI 2020 satellite event team and the MICCAI challenge reviewers for their support in the challenge workshop and review of the challenge design. We thank Kaggle for hosting the competition, providing compute resources and the competition prizes. We thank CSC - IT Center for Science, Finland, for providing computational resources. We thank Ellery Wulczyn, Annisah Um’rani, Yun Liu, and Dale Webster for their feedback on the manuscript and guidance of the project. We thank our collaborators at NMCSD, particularly Niels Olson, for internal re-use of de-identified data which contributed to the US external validation set.
4.6 The PANDA Consortium

The PANDA Challenge consortium consists of algorithm developers who participated in the PANDA challenge and pathologists who contributed to the reference standards and/or pathologist comparisons. The PANDA Challenge consortium was specifically formed for this study.

4.6.1 Pathologist participating in the international pathologists comparison

1. **Américo Brilhante**, MD, Salomão Zoppi Diagnostics/DASA, São Paulo, Brazil
2. **Aslı Çakır**, MD, Pathology Department, School of Medicine, Istanbul Medipol University, Istanbul, Turkey
3. **Xavier Farré**, MD PhD, Department of Health, Public Health Agency of Catalonia, Lleida, Catalonia, Spain
4. **Katerina Geronatsiou**, MD, Centre de Pathologie 68, Hopital Diaconat Mulhouse, Groupe Hospitalier de la Region Mulhouse Sud Alsace (GHR-MSA), Mulhouse, France
5. **Vincent Molinié**, MD PhD, Aix en Provence Hospital, Aix en Provence Hospital, Aix en Provence, France
6. **Guilherme Pereira**, MD, Histo Patologia Cirúrgica e Citologia, João Pessoa-PB, Brazil
7. **Paromita Roy**, MD, Department of Pathology, Tata Medical Center, Kolkata, India
8. **Günter Saile**, MD, Abteilung für Histopathologie und Zytologie, labor team wag, Goldach SG, Switzerland
9. **Paulo G.O. Salles**, MD PhD MBA, Belo Horizonte, Instituto Mário Penna, Brazil
10. **Ewout Schafsma**, MD PhD, Department of Pathology, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands
11. **Joëlle Tschui**, MD, Medics Pathologie, Bern, Switzerland
12. **Jorge Billoch-Lima**, MD FCAP, HRP Labs, San Juan, Puerto Rico, USA
13. **Emíio M. Pereira**, MD, Department of Pathology, Oncoclínicas group, São Paulo, São Paulo, Brazil
4.6.2 Uropathologists who set the reference standard of the US external validation set

1. Mahul B. Amin, MD, Department of Pathology and Laboratory Medicine, University of Tennessee Health Science Center, Memphis, Tennessee, USA
2. Andrew J. Evans, MD PhD, Laboratory Medicine, Mackenzie Health, Toronto, Ontario, Canada
3. Theodorus van der Kwast, MD PhD, Department of Pathology, Laboratory Medicine and Pathology, University Health Network and University of Toronto, Toronto, Ontario, Canada
4. Ming Zhou, MD PhD, Department of Pathology and Laboratory Medicine, Tufts Medical Center, Boston, Massachusetts, USA
5. Robert Allan, MD, Pathology and Laboratory Medicine Service, North Florida/South Georgia Veterans Health System, Gainesville, Florida, USA
6. Peter A. Humphrey, MD PhD, Department of Pathology, Yale School of Medicine, New Haven, Connecticut, USA

4.6.3 Algorithm developers

Overview of teams that participated in the PANDA challenge and who were accepted to join the PANDA consortium.

Dmitry A. Grechka

1. Dmitry Grechka, No affiliation, Russia

BarelyBears

1. Hiroshi Yoshihara, Department of Health Informatics, Kyoto University, Kyoto, Japan
2. Taiki Yamaguchi, Preferred Networks Inc., Tokyo, Japan
3. Kosaku Ono, Nowcast Inc., Tokyo, Japan
4. Tao Shen, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu, China

Save The Prostate

1. Rui Guo, University of Michigan, Ann Arbor, Michigan, USA
2. Chia-Lun Hsieh, No. 7, Taipei City, Taiwan
3. Igor Zubarev, No affiliation, Tula, Russia
4. **Habib S.T. Bukhar**, Janelia Research Campus, Ashburn, Virginia, USA

**PND**

1. **Yusuke Fujimoto**, Rist Inc., Tokyo, Japan
2. **Kentaro Yoshioka**, Wireless System Lab., Toshiba Corp., Kawasaki, Japan

**rahma.ai**

1. **Joni Juvonen**, Silo AI, Turku, Finland
2. **Mikko Tukiainen**, Silo AI, Turku, Finland
3. **Antti Karlsson**, University of Turku, Turku, Finland

**KovaLOVE v2**

1. **Vassili Kovalev**, PhD, Biomedical Image Analysis Department, The United Institute of Informatics Problems, Minsk, Belarus
2. **Dmitry Voynov**, Biomedical Image Analysis Department, The United Institute of Informatics Problems, Minsk, Belarus
3. **Valery Malyshev**, Biomedical Image Analysis Department, The United Institute of Informatics Problems, Minsk, Belarus
4. **Elizabeth Lapo**, Biomedical Image Analysis Department, The United Institute of Informatics Problems, Minsk, Belarus

**vanda**

1. **Kyungdoc Kim**, PhD, VUNO Inc., Seoul, Republic of Korea
2. **Byeonguk Bae**, VUNO Inc., Seoul, Republic of Korea
3. **Yeong Won Kim**, VUNO Inc., Seoul, Republic of Korea
4. **Hong-Seok Lee**, VUNO Inc., Seoul, Republic of Korea
5. **Jeonghyuk Park**, VUNO Inc., Seoul, Republic of Korea

**ctrasd123**

1. **Tianrui Chai**, School of Computer Science and Engineering, Beihang University, Beijing, China
2. **Nina Weng**, DTU Compute, Technical University of Denmark, Lyngby, Denmark

**NS Pathology**

1. **Noriaki Ota**, Systems Research & Development Center, Technology Bureau, NS Solutions Corp., Kanagawa, Japan
2. **Shinsuke Yamaoka**, Systems Research & Development Center, Technology Bureau, NS Solutions Corp., Kanagawa, Japan

Kiminya

1. **Raphael Kiminya**, No affiliation, Nairobi, Kenya

Iafoss

1. **Maxim V. Shugaev**, PhD, Department of Materials Science and Engineering, University of Virginia, Charlottesville, Virginia, USA

Aksell

1. **Shujun He**, Texas A&M University, Texas, USA
2. **Sejun Song**, No affiliation, Republic of Korea
3. **Qing Sun**, Texas A&M University, Texas, USA

ChienYiChi

1. **Jianyi Ji**, CTAccel Ltd., ShenZhen, China
2. **Arnaud Roussel**, Jumio Corp., Canada
3. **Kairong Zhou**, ELEME Inc., China

Manuel Campos

1. **Manuel Campos**, No affiliation, Madrid, Comunidad de Madrid, Spain

UCLA Computational Diagnostics Lab

1. **Wenyuan Li**, Computational Diagnostics Lab, University of California, Los Angeles, California, USA
2. **Jiayun Li**, Computational Diagnostics Lab, University of California, Los Angeles, California, USA
3. **William Speier**, Computational Diagnostics Lab, University of California, Los Angeles, California, USA
4. **Corey Arnold**, Computational Diagnostics Lab, University of California, Los Angeles, California, USA
4.7 Supplementary methods

4.7.1 Dataset inclusion and exclusion criteria

Development, tuning and internal validation set

For the development, tuning and internal validation set to be used in the PANDA challenge, we collected data from two different centers, namely Radboud University Medical Center (The Netherlands) and Karolinska Institutet (Sweden). These datasets were originally collected as part of two independent studies on automated Gleason grading.\textsuperscript{99,100} For the purpose of this challenge, the datasets were merged and further refined. We will briefly reiterate the data collection here; further details can be found in the respective papers.

For the Radboud data, we retrieved all pathology reports dated between Jan 1, 2012, and Dec 31, 2017, for patients who underwent a prostate biopsy owing to a suspicion of prostate cancer (Supplementary Figure 4.1).\textsuperscript{99} Patients were randomly sampled based on the highest reported Gleason score mentioned in each report. Additionally, a set of reports was sampled which only mentioned benign biopsies. For each patient, a single hematoxylin and eosin-stained glass slide was selected for scanning. The selected glass slides were scanned using a 3DHistech Pannoramic Flash II 250 (3DHistech, Hungary) scanner at a pixel resolution 0.24\(\mu m\). Slides were then randomly sampled to be included in the development, tuning and internal validation sets. Randomization was stratified by patient and highest Gleason pattern present in the biopsy.

The data from Karolinska comes from the Stockholm-3 diagnostic trial that was conducted between May 28, 2012 and Dec 30, 2014 (Supplementary Figure 4.2, ISRCTN84445406).\textsuperscript{100,116,117} It was a prostate cancer screening-by-invitation trial of men aged 50–69 years living in Stockholm, Sweden. The purpose of the trial was to compare prostate specific antigen (PSA) to the Stockholm-3 model (S3M) for predicting the presence of cancer, and the criterion for referral to biopsy was either PSA above 3 ng/ml or a S3M probability of 10% or higher. A single pathologist (L.E.) assessed all biopsy cores in the trial and marked out the regions of cancer next to the tissue on the glass slide with a marker pen. A random sample from the biopsies included in the trial was taken, stratified on patient and the reported Gleason score to avoid including too many of the prevalent benign and low grade diseases. The selected slides were digitized at 20X magnification using two scanners: Hamamatsu C9600-12 (Hamamatsu Photonics, Hamamatsu, Japan) and Aperio ScanScope AT2 (Leica Biosystems, Wetzlar, Germany). The pixel size at full-resolution was 0.45202\(\mu m\) (Hamamatsu) or 0.5032\(\mu m\) (Aperio). Slides were then randomly sampled to be in-
cluded in the development, tuning and internal validation sets. Randomization was stratified by patient and ISUP grade group.

To reduce the overall size of the various sets, and to achieve comparable resolution between centers, the Radboud images were downsampled and exported at a pixel spacing of 0.48\(\mu m\); the Karolinska images were exported at the original pixel spacing of 0.45\(\mu m\) or 0.50\(\mu m\) depending on the scanner. The images were exported as resolution pyramids with three levels representing downsampling factors of 1, 4 and 16 relative to the full resolution. All images were converted to TIFF format with JPEG compression and a quality setting of 70.

**US external validation set**

The US external validation set consisted of retrospective cases from three different sources, and is described in detail in a prior study (Supplementary Figure 4.3). Briefly, cases were obtained from two medical laboratories and one tertiary teaching hospital. All tumor-containing cases available from the tertiary teaching hospital from 2005-2007 were included, and a fraction of the benign biopsies available were randomly sampled for inclusion. From the medical laboratories, all available ISUP grade group 4-5 cases were included in the study, and remaining benign and ISUP grade group 1-3 cases were randomly sampled for inclusion. One representative biopsy per case was included. Biopsies with non-gradable prostate cancer variants or quality issues preventing diagnosis were excluded from the dataset. Slides were digitized on an Aperio AT2 scanner (Leica Biosystems, Wetzlar, Germany) at a resolution of 0.25\(\mu m\)/pixel (“40X magnification”). All images were converted to TIFF format with JPEG compression and a quality setting of 70.

**EU external validation set**

The EU external validation set comprised biopsy cores assessed by L.E. at the Karolinska University Hospital during 2018 (Supplementary Figure 4.4). The set included all positive biopsy cores from all men diagnosed with an ISUP grade group 2, 3, 4, or 5 cancer as well as from a random selection of men diagnosed with ISUP grade group 1 cancer during that time period. In addition, the set included all cores from a random selection of men with only benign biopsies. This resulted in 330 slides from 73 men, scanned with a Hamamatsu NanoZoomer S360 C13220-01 (Hamamatsu Photonics, Hamamatsu, Japan). The pixel size at full resolution was 0.4604\(\mu m\). All images were converted to TIFF with JPEG compression and a quality setting of 70.
4.7.2 Reference standard protocol

Development set - Radboud University Medical Center

For the cases in the development set from Radboud University Medical Center, the reference standard was determined based on the original pathology report. After scanning of the slides, trained non-experts assessed all slides and coarsely outlined each biopsy, assigning each with a Gleason score or the label ‘negative’ on the basis of the pathology report. If the pathologist report was inconclusive or lacked a description of individual biopsies, cases were flagged for a second review. If no match could be made in the second read, cases were excluded. To generate detailed label masks at gland-level, the biopsies were processed by a trained deep learning system that segmented the glands and assigned individual Gleason patterns to each gland.

Development set - Karolinska Institutet

The cases from the Karolinska Institutet that were part of the development set were reviewed by a single uropathologist (L.E.). The review of the cases was performed on the original glass slides through a microscope. The uropathologist reported both the Gleason score and the ISUP grade group for each biopsy. Additionally, the uropathologist placed pen marks on the glass slide alongside tumor tissue. Approximate label masks indicating benign and malignant tissue pixels were automatically generated based on the pen marks.

Internal validation set and tuning set - Radboud University Medical Center

The reference standard for the Radboud University Medical Center cases that were part of the tuning and validation sets was determined in three rounds. In the first round, three uropathologists (C.H.v.d.K., R.V., H.v.B.) individually graded the cases digitally using the ISUP 2014 guidelines. For a number of cases, the majority vote was taken: cases with an agreement on ISUP grade group but a difference in Gleason pattern order, e.g., 5 + 4 versus 4 + 5; cases with an equal grade group but a disagreement on Gleason score; and cases for which two pathologists agreed while the third had a maximum deviation of one grade group. Cases with a disagreement on malignancy were always flagged for a second read in round two. In the second round, each biopsy without consensus was regraded by the uropathologist whose score differed from the other two. Additional to the pathologist’s initial examination, the Gleason scores of the other pathologists were appended anonymously. Biopsies without consensus after round two were discussed in a consensus meeting.
Internal validation set and tuning set - Karolinska Institutet

The cases from the Karolinska Institutet that were part of the tuning set were reviewed by a single uropathologist (L.E.), similarly to the development set. The cases that were part of the internal validation set were initially reviewed by a single uropathologist (L.E.) on the original glass slides through a microscope. Cases initially indicated as benign were not re-reviewed. Cases indicated as malignant were divided between two other uropathologists (B.D. and H.S.), each reviewing 100 cases. In case of agreement between the first and the second review, the consensus ISUP grade group was assigned to the case. In case of disagreement, a third uropathologist (T.T.) reviewed the case. For cases that were indicated as malignant by all pathologists, the final ISUP grade group was assigned according to 2/3 consensus. If all three reviews were in disagreement, the case was excluded from the internal validation set. Any cases indicated as benign in the second or third review were excluded from the internal validation set. The second and third reviews were performed digitally using Cytomine, with all pathologists blinded to the other reviews.

US external validation set

The US external validation set was reviewed by six uropathologists (M.B.A., A.J.E., T.K., M.Z., R.A., and P.A.H.) from 6 institutions with 18 to 34 years of clinical experience after residency (mean, 25 years). Reviews were first performed by 2 of the 6 uropathologists. A third uropathologist reviewed the specimens when there were discordanices between the first two uropathologists. For cases without a majority opinion after 3 independent reviews, the median classification was used. To limit the potential ambiguity of identifying Gleason patterns due to tissue processing procedure, such as tangential cuts of the specimen, two additional adjacent sections (levels) of the specimens were also available during review. Furthermore, one additional section per specimen was stained with the PIN-4 immunohistochemistry cocktail (P504S plus p63 plus high molecular weight cytokeratin) to assist the identification of cancer tissue. The three levels and the one PIN-4 stained slide were made available to the pathologists for establishing the reference standard.

EU external validation set

The cases from the Karolinska University Hospital that were part of the EU external validation set were reviewed by a single uropathologist (L.E.). The review of the cases was performed on the original glass slides through a microscope. The uropathologist reported both the Gleason score and the ISUP grade group for each biopsy.
4.7 Supplementary methods

4.7.3 Pathologist comparison review protocol

International pathologists comparison

As part of a previous study,$^{99}$ 100 biopsies were selected to be presented to a group of pathologists in an observer experiment. Benign cases were selected manually, controlling for a broad range of tissue patterns, including inflammation and (partial) atrophy. Cases containing cancer were sampled at random, stratified for ISUP grade group. Of these 100 cases, 70 were included in the internal validation set and used for the comparison to the panel of international pathologists.

The biopsies were made available through an online viewer, PMA.view (Pathomation, Berchem, Belgium), and distributed to an external cohort of pathologists. Cohort members were invited to participate in this study at the United States and Canadian Academy of Pathology 2019 annual meeting in Washington, DC, USA (March 16–21, 2019). Interested pathologists were subsequently asked to invite colleagues in their network who had experience in Gleason grading. All pathologists who graded all biopsies were included. All cohort members had experience with Gleason grading, but to a varying degree. No time restriction was given, although we asked that they complete the grading within six weeks. In the original study, both pathologists and pathology residents were included. For the current study, only reads from pathologists were included. In total, the cohort consisted of 13 pathologists from eight countries (seven from Europe, six from outside of Europe).

US pathologists comparison

A subset of the US external validation set was reviewed by 20 US board-certified general pathologists. The pathologists reviewed the biopsies based on the 2014 ISUP grading guidelines.$^{119}$ Clinical information was not provided during grading and the pathologists were asked to review and grade biopsies as if they were reviewing a clinical slide in practice, without time constraints.

4.7.4 Available training data during the competition

For algorithm development, teams could only use the competition dataset and publicly available datasets. For public datasets, usage was only allowed if teams disclosed this beforehand on the competition’s public forum. By disclosing the use of external data, teams had no unfair advantage due to extra data availability. The use of private data, not available to other teams, was not allowed during the competition.

As part of the development set, we shared additional data besides the raw digitized biopsies to speed up development of the algorithms. As the reference standard, a
comma-separated file (CSV) was supplied that mapped each biopsy ID to a Gleason score and ISUP grade group. Additionally, each training slide had an associated label mask that contained additional information about the tissue. The label masks were generated differently per institution and contained different types of labels.

For the slides originating from Radboud, each label mask outlined the tissue within the slide. Each pixel was labeled as either background, stroma/other tissue, benign epithelium, or one of the Gleason patterns 3, 4, or 5. The label masks were generated semi-automatically using a trained deep neural network and contained label noise. Additional details on how these masks were generated can be found in the respective paper.99

For the slides originating from Karolinska Institutet, the label masks were generated based on the annotations of the pathologist who graded the development set. Each label mask outlined the tissue areas within the slide. Additionally, for slides containing cancer, areas that contained malignant tissue were coarsely outlined in the mask. Further details on how these masks were generated can be found in the respective paper.100

### 4.7.5 Kaggle competition platform

The PANDA challenge was hosted on Kaggle, one of the largest data science competition platforms. A competition typically runs for three months, during which participants or a team of participants can try to achieve the highest score in the competition’s task. A Kaggle competition consists of two leaderboards: a public leaderboard visible during the competition and a hidden private leaderboard. The public leaderboard gives teams an indication of how well their algorithm is performing during the competition. The blinded private leaderboard is used to determine the final competition ranking. Each algorithm submission is evaluated on both datasets, but only the score on the public leaderboard is shown to the teams. Because the private leaderboard is not shown, teams cannot directly tune their algorithm to score high on this leaderboard. For the PANDA challenge, the tuning set was used for the public leaderboard and the internal validation set for the private leaderboard. For the competition’s final ranking, the teams could select two of the submissions entered during the competition, of which the highest scoring one was used for the ranking.

Competitions on Kaggle often have prize money for the top-performing teams to incentivize participants to sign up and reward them for the work done. For the PANDA challenge, the top three teams on the private leaderboard were awarded monetary prizes by Kaggle.
Entering the competition on Kaggle was free and open to everyone, after agreeing to the competition rules. After signup, participants had full access to the development set, which could be downloaded directly from the Kaggle website. As part of the platform, every user had access to 30 free GPU hours per week for algorithm development. The development set was readily available when developing on the Kaggle platform, and no download was required. Additionally, participants were free to develop their algorithm offline on their hardware. Participants were asked to submit a working version of their algorithm in the form of a Jupyter Notebook or a Python script to enter the competition. This notebook or script could have associated data that contained the learned parameters of the algorithm and any other required data sources. The submitted algorithm was required to be fully self-contained, which makes it possible to reproduce the results at a later stage. The algorithm had to be developed to process all cases supplied in a specific directory automatically. After submission, the platform populated this directory with the tuning set and internal validation set cases. This processing was fully blinded to the submitter of the algorithm. Processing time was limited to 6 hours when the algorithm used a GPU. Maximum GPU memory available was 16GB. To prevent cheating, algorithms did not have internet access during this evaluation, nor could they download or upload additional data. The number of allowed submissions was limited per team and participant to a maximum of three per day. The only information disclosed to the submitter was the public leaderboard score (and not the performance on individual cases).

Through a dedicated discussion forum, participants could discuss their algorithms and problems across teams. Participants often used the discussion forum to disseminate new ideas or share additional resources. Additionally, teams could share public versions of algorithms or code snippets for others to iterate on further. One of these public notebooks was created by the organizers to kick-start the competition and showcase the dataset.

4.7.6 Methods to select the 15 teams for external validation

One month before the competition deadline, a post was placed on the competition’s discussion forum to invite teams to join the PANDA consortium. Sign-up was open to all teams that submitted a working algorithm during the competition. The deadline for signing up was July 31st, 2020, eight days after the end of the competition. When signing up, teams were asked to report: team name, team members, the data requirements of their method, whether the algorithm was based on prior work or work of other teams, and a 1000 character abstract of their method. Additionally to
the written submission, teams needed to give the challenge organizers access to their algorithm for review and further validation. In addition to the forum post, the organizers individually reached out to the top 30 teams of the competition’s leaderboard and invited them to sign up.

After the deadline, five members of the organizer team (W.B., G.L., H.P., K.K., P.R.) individually reviewed all submissions and scored them on a five-point scale (1: strong reject, to 5: strong accept). The score was based on the overall method, originality, quality of the submission, and their algorithm’s performance on the internal validation set. After scoring, all teams were discussed within the organizer team and a final ranking was established. Of all submissions, 15 teams (competition ranking in parentheses) were invited to join the PANDA consortium: PND (1), Save The Prostate (2), NS Pathology (4), Kiminya (5), BarelyBears (6), ctrasd123 (7), ChienYiChi (8), vanda (10), iafoss (11), Manuel Campos (12), Dmitry A. Grechka (18), KovaLOVE v2 (19), Aksell (20), rähmä.ai (27) and UCLA Computational Diagnostics Lab (28). Additionally, of these 15, eight teams were invited to present their method at the PANDA MICCAI workshop (October 8th, 2020, MICCAI 2020 virtual conference). All selected teams were included in the blinded validation on the external validation sets.

4.7.7 Summary of participating teams’ methods

All teams that were selected as part of the PANDA consortium were asked to report a summary of their method, including their training approach, dataset operations, and model architecture. All algorithms selected for the independent validation made use of deep learning-based techniques.

Many existing methods\(^{98}^{100}\) employ so-called patch-based training, where a biopsy is first analyzed at high magnification region by region, and the predictions for these regions are combined to determine the slide-level label. In contrast, during the competition, end-to-end training emerged as the dominant strategy for automated Gleason grading. This means that the entire WSI is treated as one data sample associated with a single target label. Due to the slides’ dimensions, a full WSI does not fit a typical GPU’s memory, and alternative methods were used to circumvent this. A popular technique was proposed by the competition participant iafoss, and adopted and modified by several participants, including the winner of the competition (PND): In brief, this approach consisted in selecting a subset of patches from a WSI based on simple filtering criteria, and processing these patches in a convolutional neural network (CNN) such that the feature representations of the patches are concatenated before they are being fed to the CNN’s classification layers. The usage of additional models to select the most informative tiles, typically operating on a low resolution...
version of the WSI, was also proposed (Aksell, ChienYiChi, Save The Prostate, UCLA Computational Diagnostics Lab). Other training approaches included for example using CNNs as patch-wise feature extractors, followed by a recurrent neural network for aggregating the patch-level feature representation to a WSI-level output (Dmitry A. Grechka).

Besides the use of end-to-end training, a second common denominator among leading competition participants was the extensive use of data cleaning. The development set contained substantial levels of label noise. For the cases of Radboud University Medical Center, the labels were generated semi-automatically. For the cases from Karolinska Institute, the reference standard was based on the assessment of a single pathologist, and the semi-automatically generated pixel-level labels indicated malignant regions only in an approximate manner. Several teams used (semi-)automatic techniques to remove slides with artifacts, low-quality slides, and images where teams expected errors in the reference standard. For example, BarelyBears used Online Uncertainty Sample Mining and excluded 10% of training patches associated with the highest average loss values during training. PND and Dmitry A. Grechka identified samples with potentially erroneous labels after completing model training, based on the predicted ISUP grade group differing from the target label by more than a specified amount. Training was then repeated after exclusion of such samples. In the approaches of iafoss and Save The Prostate, the training labels were adjusted iteratively instead of excluding samples.

Interestingly, while some teams experimented with multi-resolution approaches and used low resolution images for selecting regions of interest, virtually all of the proposed solutions used the intermediate resolution level of the input WSI for producing the final ISUP grade group predictions. This corresponds to approximately 2 μm pixel spacing (and a typical magnification of 5X), which can be considered a relatively coarse resolution.

Finally, to improve performance and increase the algorithms’ generalization ability, all of the 15 teams utilized ensembles of multiple models. Ensemble strategies differed from team to team, ranging from models trained using different hyperparameters, different patch selection strategies or different loss functions to a set of different neural network architectures combined into an ensemble.
4.8 Supplementary figures

Supplementary Figure 4.1: Flow chart of inclusion and exclusion for data originating from Radboud University Medical Center (development, tuning and internal validation sets, and international pathologist comparison).
Supplementary Figure 4.2: Flow chart of inclusion and exclusion for data originating from Karolinska Institutet (development, tuning and internal validation sets).
Supplementary Figure 4.3: Flow chart of inclusion and exclusion for data originating from the United States (US external validation set, and US pathologist comparison).
Supplementary Figure 4.4: Flow chart of inclusion and exclusion for data originating from Karolinska University Hospital (EU external validation set).
Supplementary Figure 4.5: Individual algorithms’ agreement with the reference standard for the validation sets.
Supplementary Figure 4.6: Individual algorithms’ sensitivity and specificity for the validation sets.
Supplementary Figure 4.7: Algorithms' predictions for the internal validation set.
Supplementary Figure 4.8: Algorithms’ predictions for the US external validation set.
**Supplementary Figure 4.9**: Algorithms’ predictions for the EU external validation set.
### 4.9 Supplementary tables

**Supplementary Table 4.1:** Scanner details for the slides in the development and tuning set.

<table>
<thead>
<tr>
<th>Source</th>
<th>EU Development set</th>
<th>EU Tuning Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Radboud University Medical Center, Netherlands</td>
<td>Radboud University Medical Center, Netherlands</td>
</tr>
<tr>
<td></td>
<td>Karolinska Institutet, Sweden</td>
<td>Karolinska Institutet, Sweden</td>
</tr>
<tr>
<td>Scanning equipments</td>
<td>3DHistech Pannoramic Flash II 250 (3DHistech, Hungary)</td>
<td>Aperio AT2 (Leica, Germany) &amp; Hamamatsu C9600-12 (Hamamatsu, Japan)</td>
</tr>
<tr>
<td></td>
<td>3DHistech Pannoramic Flash II 250 (3DHistech, Hungary)</td>
<td>Aperio AT2 (Leica, Germany) &amp; Hamamatsu C9600-12 (Hamamatsu, Japan)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pixel spacing of original scanned slides</th>
<th>EU Development set</th>
<th>EU Tuning Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.24 µm and 0.45 µm</td>
<td>0.50 µm and 0.45 µm</td>
<td>0.50 µm and 0.45 µm</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pixel spacing of downsampled slides, available to the algorithms</th>
<th>EU Development set</th>
<th>EU Tuning Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.48 µm and 0.45 µm</td>
<td>0.50 µm and 0.45 µm</td>
<td>0.50 µm and 0.45 µm</td>
</tr>
</tbody>
</table>
**Supplementary Table 4.2:** Scanner details for the slides in the internal and external validation sets.

<table>
<thead>
<tr>
<th>Source</th>
<th>EU Internal Validation Set</th>
<th>US External Validation set</th>
<th>EU External Validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Radboud University Medical Center, Netherlands</td>
<td>Karolinska Institutet, Sweden</td>
<td>Medical Laboratories, CA/UT, USA; Tertiary Teaching Hospital, CA, USA</td>
</tr>
<tr>
<td></td>
<td>Karolinska University Hospital, Sweden</td>
<td></td>
<td>Karolinska University Hospital, Sweden</td>
</tr>
<tr>
<td>Scanning equipments</td>
<td>3DHistech Pannoramic Flash II 250 (3DHistech, Hungary)</td>
<td>Aperio AT2 (Leica, Germany) &amp; Hamamatsu C9600-12 (Hamamatsu, Japan)</td>
<td>Aperio AT2 (Leica, Germany)</td>
</tr>
<tr>
<td></td>
<td>Hamamatsu C13220-01 (Hamamatsu, Japan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pixel spacing of original scanned slides</td>
<td>0.24 µm</td>
<td>0.50 µm and 0.45 µm</td>
<td>0.25 µm</td>
</tr>
<tr>
<td>Pixel spacing of downsampled slides, available to the algorithms</td>
<td>0.48 µm</td>
<td>0.50 µm and 0.45 µm</td>
<td>0.50 µm</td>
</tr>
</tbody>
</table>
Supplementary Table 4.3: Ensemble algorithm performance. An ensemble was created by taking the majority vote for each case in the validation sets and computing the agreement of this ensemble with the reference standards.

<table>
<thead>
<tr>
<th>Dataset / Metric</th>
<th>Internal validation set</th>
<th>US external validation set</th>
<th>EU external validation set</th>
<th>International pathologists comparison</th>
<th>US pathologists comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratically weighted kappa (95% CI)</td>
<td>0.940 (0.928-0.952)</td>
<td>0.892 (0.874-0.909)</td>
<td>0.899 (0.869-0.924)</td>
<td>0.887 (0.820-0.934)</td>
<td>0.870 (0.826-0.904)</td>
</tr>
<tr>
<td>Sensitivity (95% CI)</td>
<td>99.7% (99.0-100.0)</td>
<td>99.2% (98.3-99.8)</td>
<td>98.2% (96.3-99.6)</td>
<td>100.0% (100.0-100.0)</td>
<td>99.0% (97.4-100.0)</td>
</tr>
<tr>
<td>Specificity (95% CI)</td>
<td>96.4% (93.8-98.7)</td>
<td>81.9% (77.1-86.6)</td>
<td>89.8% (83.8-95.1)</td>
<td>100.0% (100.0-100.0)</td>
<td>82.5% (69.7-93.5)</td>
</tr>
</tbody>
</table>
AI-assisted Gleason grading

Authors: Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Aslı Çakır, Lars Egevad, Martin Eklund, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, Guilherme Pereira, Paromita Roy, Günter Saile, Paulo Salles, Ewout Schaafsma, Joëlle Tschui, Anne-Marie Vos, ISUP Pathology Imagebase Expert Panel, Hester van Boven, Robert Vink, Jeroen van der Laak, Christina Hulsbergen-van der Kaa & Geert Litjens

Original title: Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists

Published in: Modern Pathology (Volume: 34, issue: 3, March 2021)

DOI URL: doi.org/10.1038/s41379-020-0640-y
Abstract

The Gleason score is the most important prognostic marker for prostate cancer patients, but it suffers from significant observer variability. Artificial intelligence (AI) systems based on deep learning can achieve pathologist-level performance at Gleason grading. However, the performance of such systems can degrade in the presence of artifacts, foreign tissue, or other anomalies. Pathologists integrating their expertise with feedback from an AI system could result in a synergy that outperforms both the individual pathologist and the system. Despite the hype around AI assistance, existing literature on this topic within the pathology domain is limited. We investigated the value of AI assistance for grading prostate biopsies. A panel of 14 observers graded 160 biopsies with and without AI assistance. Using AI, the agreement of the panel with an expert reference standard increased significantly (quadratically weighted Cohen's kappa, 0.799 vs. 0.872; p = 0.019). On an external validation set of 87 cases, the panel showed a significant increase in agreement with a panel of international experts in prostate pathology (quadratically weighted Cohen's kappa, 0.733 vs. 0.786; p = 0.003). In both experiments, on a group-level, AI-assisted pathologists outperformed the unassisted pathologists and the standalone AI system. Our results show the potential of AI systems for Gleason grading, but more importantly, show the benefits of pathologist-AI synergy.
5.1 Introduction

The biopsy Gleason score is the most important tissue-based prognostic marker for prostate cancer patients. However, it has been shown that Gleason grading suffers from significant inter- and intraobserver variability. Specialized uropathologists show higher concordance rates, but such expertise is not always available. Artificial intelligence (AI) systems based on deep learning have achieved pathologist-level performance in Gleason grading, but it is not yet investigated whether pathologists improve in Gleason grading if they are assisted by such systems.

Pathologists assess the Gleason grade of a prostate biopsy through microscopic assessment of tissue stained with hematoxylin and eosin (H&E). Based on the morphological pattern of the tumor, a grade between one and five is assigned, with one being the lowest and five the highest. For biopsies, the Gleason score is the sum of the two most common patterns, e.g., $3 + 5 = 8$. If a higher tertiary pattern is present, this is used instead of the secondary pattern. Patterns 1 and 2 are not reported anymore for biopsies.

Recently, (ISUP) grade groups were introduced which aimed to improve the reporting of Gleason grading by assigning the Gleason score to one of five prognostic groups. These groups are directly based on the Gleason score; $3 + 3$ and lower go to group 1, $3 + 4$ to group 2, $4 + 3$ to group 3, $3 + 5$, $5 + 3$ and $4 + 4$ to group 4, and higher scores to group 5. While the introduction of grade groups showed clinical value and increased interpretability of the tumor grade for patients, it has not improved the inter- and intraobserver variability.

Deep learning has shown promise in many medical fields, and the introduction of digital pathology allows for AI-based diagnostics in pathology. For prostate cancer, methods based on deep learning have been developed for tumor detection, grading of prostatectomies, tissue microarrays, and biopsies. In multiple studies, such deep learning systems showed pathologist-level performance, within the limits of the study setup.

Although deep learning systems have shown to achieve high performances on grading tasks, evidence of the merit of such systems when embedded in the pathologist's workflow is limited. Deep learning systems can be viewed as a new tool for pathologists to use in their diagnostic process and should also be evaluated as such. In addition, regardless of the merits, most developed systems are also constrained by significant limitations that affect the performance and can lower the diagnostic power. Within histopathology, the presence of non-prostate tissue, atypical tissue patterns, ink on a slide, fixation, scanning and cutting artifacts, or the presence of rare cancer sub-
types can dramatically affect a system’s assessment of tissue. Many of these errors, especially those caused by artifacts, are easily spotted by a human observer. Studies combining experts’ opinions with feedback from automated systems have mainly been performed outside of the field of pathology; for example on the task of breast cancer detection in mammography.\textsuperscript{121} For pathology, on the task of cancer metastasis detection in lymph nodes, the sensitivity of detection of micrometastases increased, and overall case reading time went down as a result of AI support.\textsuperscript{122} On the task of mitosis counting, AI-generated hotspots improved reproducibility between readers.\textsuperscript{123} For prostate cancer, AI assistance has shown potential in increasing sensitivity for detecting cancer in biopsies.\textsuperscript{120} However, most of these studies focus either on computer-aided detection or diagnosis. For prognostic measures, such as Gleason grading of prostate biopsies, there is, to the best of our knowledge, no such study as of yet.

In a previous study, we developed a fully automated deep learning system for grading prostate cancer.\textsuperscript{99} The deep learning system was trained on a large dataset of prostate biopsies and achieved pathologist-level performance, both in determining the grade group and in stratifying patients in relevant risk categories. As part of the initial validation of the system, its performance was compared with a panel of pathologists in an observer experiment. The deep learning system outperformed 10 out of 15 observers on determining the grade group.

In this study, we investigate the value of AI-assisted reading by pathologists for Gleason grading of prostate biopsies by comparing the diagnostic performance of pathologists with and without the assistance of a deep learning system.

\section{Materials and methods}

\subsection{Collection of the dataset and setting the reference standard}

In a previous study,\textsuperscript{99} we developed a deep learning system to grade prostate biopsies using the Gleason grading system. To train this system, we collected a dataset of 5759 H&E-stained biopsies from 1243 patients. All biopsy procedures were performed as part of routine diagnostics at the Radboud University Medical Center between 2012 and 2017. For the study, the H&E-stained glass slides of the biopsies were digitized at $20\times$ magnification (pixel resolution $0.24\mu m$) using a 3DHistech Pannoramic Flash II 250 scanner and subsequently anonymized. The need for informed consent was waived by the local ethics review board (IRB number 2016–2275).

Of the dataset, 550 biopsies were excluded from model development and used as an independent test set to evaluate the deep learning system. Patients that were included
in this test set were independent of the patients in the training set. Given the inter-
observer variability of Gleason grading, validation of the system required a consensus
reference standard. In the first round, three expert pathologists (CH-vdK, RV, HvB)
with a subspecialty in uropathology individually graded the cases in the test set using
the ISUP 2014 guidelines.\(^{19}\) For some cases the majority vote was taken: cases with
an agreement on grade group but a difference in Gleason pattern order, e.g., 5 + 4
versus 4 + 5; cases with an equal grade group but a disagreement on Gleason score;
and cases for which two pathologists agreed while the third had a maximum deviation
of one grade group. Cases with a disagreement on malignancy were always flagged.
In the second round, cases that had no agreement were presented to the pathologist
who deviated the most from the other two. Additional to the pathologist’s score,
the scores of the two other pathologists were shown anonymously. Finally, biopsies
without agreement after two rounds were discussed in a consensus meeting.

5.2.2 Observer panel and case selection

Part of the first study was a comparison of the deep learning system to a panel of
pathologists. Of the full test set, 100 cases were selected and presented to a panel
of 13 external pathologists and two pathologists in training. Of these 100 cases, 20
benign cases were selected by one of the expert pathologists (CH-vdK). The benign
cases were chosen to cover the full spectrum of possible pitfalls for cancer, including
partial atrophy, reactive atypia, granulomatous inflammation with epithelioid cells,
atypical adenomatous hyperplasia as well as HGPIN (Table 5.1). The other 80 cases
were sampled based on the grade group assignment by the same pathologist, selecting
an equal number of cases per grade group. Potential pitfalls in the set of malignant
cases are shown in Table 5.2. The panel was asked to grade all biopsies through an
online viewer PMA.view (Pathomation, Berchem, Belgium) following the ISUP 2014
guidelines. No time limit was set for the grading process.

All pathologists that participated in the first study were invited to participate in the
present study. One additional pathologist in training, who showed interest in the first
study but was not able to grade all biopsies before submission of the previous paper,
was also asked to join the current study. Panel members were not involved in the
development of the AI system, nor had used the system before this study.

We included all 100 biopsies from the first study, as the panel already graded these
cases. In addition, we extended the dataset with 60 new cases from the original
test set, all of which were unseen by the panel members. These new unseen cases
were used as control cases to measure the potential effect of a second-read on the
original cases. One of the expert pathologists who set the reference standard (CH-
Table 5.1: Description and presence of inflammation for the benign cases of the internal test set.

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Description</th>
<th>Inflammation</th>
<th>Case ID</th>
<th>Description</th>
<th>Inflammation</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>AAH</td>
<td>Mild</td>
<td>109</td>
<td>RA + BCH + A</td>
<td>Mild</td>
</tr>
<tr>
<td>21</td>
<td>PA</td>
<td>Mild</td>
<td>113</td>
<td>none</td>
<td>None</td>
</tr>
<tr>
<td>23</td>
<td>RA + A</td>
<td>Mild</td>
<td>171</td>
<td>PA + A</td>
<td>None</td>
</tr>
<tr>
<td>30</td>
<td>RA + BCH + A</td>
<td>Mild</td>
<td>192</td>
<td>RA + A</td>
<td>Mild</td>
</tr>
<tr>
<td>33</td>
<td>HGPIN + RA + A</td>
<td>Mild</td>
<td>227</td>
<td>PA + HGPIN</td>
<td>Minimal</td>
</tr>
<tr>
<td>36</td>
<td>RA</td>
<td>Minimal</td>
<td>249</td>
<td>RA</td>
<td>Moderate</td>
</tr>
<tr>
<td>38</td>
<td>RA + A</td>
<td>Mild</td>
<td>280</td>
<td>A + PA</td>
<td>Minimal</td>
</tr>
<tr>
<td>66</td>
<td>PA</td>
<td>Minimal</td>
<td>284</td>
<td>RA + PA</td>
<td>Mild</td>
</tr>
<tr>
<td>67</td>
<td>AAH</td>
<td>None</td>
<td>287</td>
<td>HGPIN</td>
<td>None</td>
</tr>
<tr>
<td>68</td>
<td>A</td>
<td>None</td>
<td>326</td>
<td>PA + AAH</td>
<td>None</td>
</tr>
<tr>
<td>82</td>
<td>None</td>
<td>Minimal</td>
<td>333</td>
<td>A</td>
<td>None</td>
</tr>
<tr>
<td>88</td>
<td>RA</td>
<td>Moderate Gr</td>
<td>348</td>
<td>HGPIN + RA + PA</td>
<td>Minimal</td>
</tr>
<tr>
<td>90</td>
<td>RA</td>
<td>Severe Gr</td>
<td>398</td>
<td>RA + PA + BCH</td>
<td>Mild</td>
</tr>
<tr>
<td>94</td>
<td>PA</td>
<td>Mild</td>
<td>430</td>
<td>RA</td>
<td>Moderate</td>
</tr>
<tr>
<td>108</td>
<td>PA</td>
<td>Mild</td>
<td>482</td>
<td>PA</td>
<td>None</td>
</tr>
</tbody>
</table>

Abbreviations: A (full) atrophy, AAH atypical adenomatous hyperplasia, PA partial atrophy, RA reactive atypia, BCH basal cell hyperplasia, HGPIN high grade PIN, Gr granulomatous.

vdK), selected ten benign cases manually, again controlling for a broad range of tissue patterns. The remaining fifty cases were sampled based on the consensus grade group, selecting an equal number of cases per group. All 160 biopsies were shuffled and assigned new identifiers. This dataset is further referenced as the internal dataset.

5.2.3 Feedback of the AI system

We processed each biopsy in the dataset using the deep learning system, resulting in a prediction of the volume percentages of each Gleason pattern (if present), the Gleason score, and the grade group per biopsy. Besides a numerical prediction, the system also generated an overlay that outlined malignant glands: Gleason pattern 3 in yellow, Gleason pattern 4 in orange, and Gleason pattern 5 in red. For the current experiment, we chose not to highlight detected benign tissue. The overlays were postprocessed by a connected components algorithm to remove small artifacts and to ensure that each detected malignant gland was assigned to a single Gleason pattern. Postprocessing was done fully automatically without manual review. All
5.2 Materials and methods

Table 5.2: Potential pitfalls in the set of malignant cases of the internal test set.

<table>
<thead>
<tr>
<th>Case ID</th>
<th>Comment</th>
<th>Case ID</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>IDC dd HGPIN and G4 cribriform</td>
<td>277</td>
<td>IDC dd G4 cribriform</td>
</tr>
<tr>
<td>14</td>
<td>G5 dd -itis 3+</td>
<td>317</td>
<td>IDC dd HGPIN and G4 cribriform</td>
</tr>
<tr>
<td>112</td>
<td>G5 dd -itis 3+</td>
<td>370</td>
<td>IDC dd HGPIN and G4 cribriform</td>
</tr>
<tr>
<td>114</td>
<td>G5 dd -itis 3+</td>
<td>391</td>
<td>Partly foamy gland, no dd problem</td>
</tr>
<tr>
<td>132</td>
<td>Minimal cancer</td>
<td>395</td>
<td>Partly foamy gland, no dd problem</td>
</tr>
<tr>
<td>141</td>
<td>HGPIN dd G3</td>
<td>396</td>
<td>Partly foamy gland, no dd problem</td>
</tr>
<tr>
<td>186</td>
<td>Partly foamy gland, no dd problem</td>
<td>424</td>
<td>Foamy gland and IDC dd invasive</td>
</tr>
<tr>
<td>191</td>
<td>Partly foamy gland, no dd problem</td>
<td>439</td>
<td>IDC dd invasive G4</td>
</tr>
<tr>
<td>194</td>
<td>IDC dd HGPIN and G4 cribriform</td>
<td>448</td>
<td>Hyperplastic variant</td>
</tr>
<tr>
<td>212</td>
<td>Partly atrophic subtype</td>
<td>468</td>
<td>dd HGPIN vs. invasive</td>
</tr>
<tr>
<td>216</td>
<td>Foamy subtype</td>
<td>475</td>
<td>IDC dd HGPIN and G4 cribriform</td>
</tr>
<tr>
<td>229</td>
<td>IDC dd HGPIN and G4 cribriform</td>
<td>476</td>
<td>IDC dd G4 cribriform</td>
</tr>
<tr>
<td>234</td>
<td>Minimal cancer</td>
<td>526</td>
<td>IDC dd invasive G4</td>
</tr>
<tr>
<td>241</td>
<td>Minimal cancer and G5 dd -itis</td>
<td>529</td>
<td>HGPIN dd invasive G3</td>
</tr>
<tr>
<td>242</td>
<td>Partly hyperplastic type</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: IDC intraductal carcinoma; HGPIN high grade PIN; G3, G4, G5 growth pattern 3, 4, 5; dd differential diagnosis; -itis inflammation.

selected biopsies were used, regardless of the correctness of the system’s prediction on these biopsies.

5.2.4 Second-read with AI assistance

The 160 biopsies were made available to the panel of pathologists through the same online viewer as during the first read. The time between the first and second read was at least 3 months. Each biopsy could be viewed at a maximum pixel spacing of $0.24 \mu m$ (roughly equivalent to $40 \times$ objective magnification). Next to the original biopsy, we showed an exact copy of the biopsy where AI-predicted Gleason patterns were highlighted using different colors (Figure 5.1). This overlay could be used to assess the tissue that the algorithm had flagged as malignant. To complement the overlay, we also supplied the numerical output of the deep learning system to the panel, including the predicted volume percentages, the presence of tumor (yes/no), the Gleason score, and the grade group.

We asked each panel member to report: whether a biopsy contained tumor, the presence of Gleason patterns, volume percentages of present patterns, and the grade
Figure 5.1: Overview of the viewer used in the observer experiment. Both the original biopsy (a) and the biopsy with the AI overlay (b) are presented to the pathologist. Each individual tumor gland is marked by the deep learning system in the overlay. The case-level grade group was supplied to the panel as part of their (separate) grading form.

Group. After grading a case, the panel members had to indicate whether they thought the system’s prediction influenced their assessment.

No time restriction was given per case, but we asked each pathologist to complete all 160 cases within 8 weeks. Each panel member was instructed to review the cases individually without consulting colleagues. Panel members had no access to the cases from the previous experiment nor to the grades that they assigned. Between the first and second read, no performance indication or feedback was given to panel members with respect to the reference standard. When all cases were graded, each panel member was asked to fill in a questionnaire regarding the process and feedback from the deep learning system.

5.2.5 External dataset

After the experiment on the internal dataset, we performed an additional experiment to test our hypothesis on external data. The time between the main experiment, and this external validation was 6 months. For this external validation, we made use of the Imagebase dataset. The Imagebase set consists of 90 cases of prostate needle biopsies with cancer independently graded by 24 international experts in prostate
5.2 Materials and methods

pathology. Grading by the experts was done between May and September 2015, based on microphotographs taken from representative locations.

Each glass slide in the dataset consisted of two sections of the same biopsy, of which one was marked by pen. The biopsies were scanned on two different scanners as part of a previous independent study on automated Gleason grading by Ström et al.\textsuperscript{100} We extracted the marked biopsies and removed the pen markings. The deep learning system was applied as-is to the dataset without any normalization of the data. In the absence of a training set, the decision thresholds of the system were not optimized on the dataset. Instead, we determined that any detected tumor would classify the biopsy as malignant and used a 5% volume threshold for the inclusion of secondary patterns, comparable with clinical practice.

The setup of the experiment on the external dataset was equal to the experiment on the internal dataset. All pathologists who took part in the first experiment were invited to join the second experiment. Pathologists were given 1 week to grade the Imagebase cases. Instead of using microphotographs, pathologists were given access to the full biopsy through the digital viewer. After the unassisted read and a 2-week washout-period, the pathologists had to reexamine the cases with AI assistance.

5.2.6 Statistical analysis

After all panel members completed the grading of the biopsies, we compared their raw scores to the consensus reference standard. Scores were given on a six-point scale: 0 for benign, and 1–5 for grade groups. Cohen’s kappa with quadratic weights was used as the primary metric of performance. On a group-level, we used the median kappa as the metric to account for outliers.

To compare reading cases with or without AI assistance we conducted a statistical analysis, using the difference in kappa between the two reads as the test statistic. A Shapiro–Wilks test for normality was performed to show that the data were not normally distributed. To compare the difference in kappa scores, we performed a Wilcoxon signed-rank test on the paired kappa values. The test statistic was computed using the grades of the 100 cases that were used in both reads.

To account for possible bias in the reference standard, we computed the pairwise agreement between all panel members individually. The reference standard was not used in this analysis. Agreement was calculated using quadratically weighted Cohen’s kappa on the grade group.

In addition to a comparison of grade group agreement, we compared the concordance on estimated tumor volume. For each panel member, we computed Pearson’s correlation on the reported tumor volume with all other panel members (pairwise
combinations). Correlations were computed on data of the 80 malignant cases that were used in both reads.

For the external Imagebase dataset, we used agreement using quadratically weighted Cohen’s kappa as the main metric, to allow for a comparison between the internal and external data. Agreement was calculated using linear weights in the Imagebase study, so we additionally computed the agreement using linear weights to compare between the two studies. As there was no consensus between Imagebase panel members for every case, we computed the average agreement of each panel member of the study pairwise with the Imagebase panel members, following similar work on this dataset. To compare between reads, the median value of the pairwise kappas was used. A Shapiro–Wilks test for normality was performed to show that the data were not normally distributed. To compare the difference in kappa scores, we performed a Wilcoxon signed-rank test on the paired kappa values.

All statistical analyses were performed using Python (version 3.7.6), the pandas package (version 1.0.1), the scikit-learn package (version 0.22.2) and the SciPy package (version 1.4.1). Figures were generated using the matplotlib package (version 3.1.3).

5.3 Results

5.3.1 The observer panel

We invited 16 pathologists (board certified or residents) who participated in an earlier study on automated Gleason grading to perform this observer experiment. Two panel members dropped out due to other obligations or a lack of time. In total, the observer panel consisted of 14 members (11 certified pathologists and 3 pathology residents), originating from 12 independent labs and 8 countries. All panel members had prior experience with Gleason grading, though with varying amounts of experience.

5.3.2 Dataset under review and reference standard

From the test set that was previously used to evaluate our deep learning system, a set of 160 cases was selected to be reviewed by the panel. All cases under review had been graded by three uropathologists with extensive (>20 years) experience in Gleason grading, and their consensus opinion set the reference standard. The agreement between the uropathologists in the first round of the consensus-protocol was high (quadratic weighted Cohen’s kappa 0.925).
5.3 Results

Of the selected cases, 100 cases were already graded by the panel as part of the previous study and were reused for the current study; the remaining 60 cases were unseen to act as controls, to measure the potential effect of a second-read on the original cases. The complete set of 160 cases for the AI-assisted read in the present study consisted of 30 (19%) benign cases, 22 (14%) cases with grade group 1, 26 (16%) cases with grade group 2, 32 (20%) cases with grade group 3, 20 (13%) cases with grade group 4 and 30 (19%) cases with grade group 5.

5.3.3 AI-assisted Gleason grading

After grading, all panel members filled in a questionnaire on the grading process. Five out of 14 (36%) panel members predicted that they scored somewhat better in comparison with the first read. Of these five, 3 out of 14 (21%) expected a performance increase due to being more experienced in viewing cases using the online viewer, and 2 out of 14 (14%) because of the AI assistance. The majority of the panel members (8 out of 14, 57%) indicated they did not expect a performance increase as a result of the AI assistance, while one pathologist (1 out of 14, 7%) expected to have scored somewhat lower.

Eleven out of 14 (79%) panel members indicated that they used feedback from the AI system during grading. Of all the components of the AI feedback, the Gleason pattern overlay was determined to be the most useful and easy to interpret (Figure 5.2). The panel members indicated that the final grade group, as assigned by the system, was the least helpful. The majority of panel members noted that the AI assistance did not distract them from the grading process, but instead made grading the biopsies faster (Figure 5.3).

5.3.4 Performance of the panel with and without AI feedback

In the first read without AI assistance, the agreement with the reference standard (measured by the median quadratically weighted Cohen’s kappa) for the panel was 0.799. In the second, AI-assisted read, the median kappa of the panel increased to 0.872 (9.14% increase), showing a significant increase in performance (Wilcoxon signed-rank test p = 0.019, Figure 5.4). On the same dataset, the AI system in itself achieved a kappa score of 0.854. Excluding panel members who estimated that they improved due to viewing more cases (n = 3) or excluding pathologists who indicated that they did not use the AI feedback (n = 3), we found a comparable increase in median kappa from, respectively 0.754–0.875 (p = 0.041) and 0.754–0.870 (p = 0.016).
Figure 5.2: Survey results on the AI feedback. Panel members were asked to indicate how useful each part of the AI’s feedback was on a five-point scale from “Not useful” to “Very useful”.

Figure 5.3: Survey results on the grading process. Panel members were asked to reflect on the grading process and answer questions on a five-point scale from “Strongly disagree” to “Strongly agree”.
5.3 Results

Figure 5.4: Panel performance with and without AI assistance. With AI assistance, the median performance of the group increased while the variability between panel members went down.

Nine of the 14 (64%) panel members scored higher in the assisted read, while five (36%) panel members scored slightly lower, though with a maximum decrease in kappa score of 0.013. Of the five that scored lower, four already outperformed the AI system in the first read. The interquartile range of the panel’s kappa values dropped from 0.113 to 0.073 in the second read (Figure 5.4).

In the first read, the kappa value of the AI system exceeded that of 10 out of the 14 (71%) panel members. In the AI-assisted read, only five of the panel members (36%) scored a kappa value below that of the AI system. The largest improvement was seen for panel members who had less than 15 years of experience (Figure 5.5a). Of the panel members who scored lower than the AI system in the unassisted read (10 out of 14, 71%), nine scored higher in the assisted read (9 out of 10, 90%). None of the panel members who outperformed the AI in the unassisted read improved in the assisted read. On a group-level, the median performance of AI-assisted reads was higher than both that of the standalone AI system and the unassisted reads.

The agreement of the panel with the reference standard on the control cases was high, with a median kappa value of 0.910, and slightly higher in comparison with the test cases (kappa 0.872). The control cases were only viewed in the assisted read. The system’s performance on the control cases was also higher, with a kappa value of 0.905 compared with 0.854 on the test cases.
Figure 5.5: Individual performance of panel members shown for both the unassisted read (light blue) and assisted read (dark blue). Results for the internal test set shown in (a) and external test set shown in (b). Lower performance in the unassisted read is indicated with a line in the light blue bars. Pathologists are sorted based on experience level and the kappa value of the unassisted read. The performance of the standalone AI system is shown in green. In the unassisted reads, the AI system outperforms the group. In the assisted reads, the median performance of the group is higher than of the standalone AI system.
Between panel members, grading became more consistent in the assisted read. In the unassisted read, without taking the consensus into account, the median pair-wised agreement within the panel was 0.737 (quadratically weighted Cohen’s kappa). In the assisted read, this agreement increased to 0.859 (Figure 5.6).

One pathologist did not report total tumor volume in the unassisted read and was excluded for the analysis of reported tumor volume. The correlation between the remaining panel members (13 out of 14) on total tumor volume was high in the unassisted read (mean Pearson’s $r = 0.744$), and increased slightly the assisted read ($r = 0.780$). Variation between panel members was lower in the assisted read; the interquartile range decreased from 0.164 to 0.105 (Figure 5.7).

### 5.3.5 External validation

For the Imagebase experiment, three cases and scores from one of the expert pathologists were excluded following the procedure by Ström et al.\textsuperscript{100} The remaining 87 cases all contained tumor, and the Imagebase panel had an average pairwise Cohen's kappa of 0.819 (quadratically weighted) and 0.677 (linear weighted). The deep learn-
Figure 5.7: Pairwise correlations on reported total tumor volume between panel members. While only a slight increase can be observed in the assisted read, the total variation dropped substantially.
5.4 Discussion

To the best of our knowledge, this study was the first to explore the possible merits of AI assistance on histological tumor grading. In a research setting, we showed that AI assistance improves pathologists’ performance at Gleason grading of prostate biopsies. Measured through the agreement with an expert reference standard, the read with AI assistance resulted in a significant increase in performance on the internal test set (quadratically weighted Cohen’s kappa, 0.799 vs. 0.872).

On an external validation set, Imagebase, the same positive effect of AI assistance was shown. With respect to a panel of international experts in prostate pathology, agreement increased from 0.733 to 0.786 (quadratically weighted Cohen’s kappa). In comparison with the internal set, the panel and AI system both scored lower on this external set. This could be explained due to several reasons: the Imagebase cases were collected specifically to represent a wide range of tissue patterns and the panel who set the reference standard reached consensus in only 50 of the cases, showing significant difficulty in the cases. Second, some differences can be accounted to the reference standard being set in 2015 and the use of microphotographs instead of whole-slide images by the expert panel. Last, the external dataset was collected in a different lab and scanned using a different scanner, which could negatively influence the accuracy of the feedback provided by the AI system.

Variance between panel members’ performance decreased due to AI assistance, resulting in overall more consistent grading. This decrease in grader variability was observed in comparison with the reference standard, and between panel members on both grade group and tumor volume estimation. Reduced observer variability of Glea-
AI-assisted Gleason grading is highly desirable, as it could lead to a stronger prognostic marker for individual patients and reduces the effect of the diagnosing pathologist on potential treatment decisions.

In the unassisted read, the AI system outperformed 10 out of 14 pathologists, and this dropped to only 5 out of 14 in the second-read with AI assistance. Pathologists assisted by the AI system not only improved compared with unassisted reads but also achieved higher median performance than the standalone AI. These results indicate that there is a potential benefit of pathologists using AI assistance as a supportive tool during diagnosis. Especially in geographic regions where the number of pathologists is limited or subspecialized pathologists are not available, AI systems such as ours can support pathologists in achieving higher grading accuracy and consistency.

The most substantial increase in performance was seen for panel members who initially scored lower than the AI system. Most of the pathologists with more than 15 years of experience, who often outperformed the AI system in the unassisted read, scored comparably in both reads. Some pathologists’ scores approached the agreement between the pathologists who set the reference standard. In such cases, given the subjective nature of Gleason grading, objective improvement is difficult to determine. In the external set, which had a higher case difficulty, AI assistance improved the scores of all but one of the panel members.

While no performance gain was found for some pathologists in terms of diagnostic accuracy, most pathologists indicated that the use of the AI system led to faster grading. However, in this study, we did not directly measure the time taken per case nor did we limit the maximum time per case. For clinical applications, where reducing the workload and overall efficiency is an important topic, saving time through AI assistance is of great interest. Additional research could quantitively test the ability of AI systems to reduce time needed per case.

Through a questionnaire, we investigated the pathologists’ experiences when using the system. One of the design goals of the deep learning system was to support the workflow of pathologists. The system was developed to give feedback on multiple abstraction levels, with the grade group giving an overall assessment and the overlay more detailed feedback. We assumed that the precise gland-level segmentations of the tumor and Gleason patterns could support pathologists in quickly assessing glandular regions and assisting in volume measurements. Almost all pathologists indicated that the AI system’s overlay was useful, and, based on the questionnaire, was the most used feature of the system. AI assistance through these overlays can be seen as another tool for pathologists that gives feedback on a glandular level, comparable with, e.g., immunohistochemistry, and gives direct support for the systems’ case-level
prediction. The overlays allow pathologists to combine their expertise with the added feedback of the system to determine the final grade.

The AI feedback was also given on a case-level through volume measurements, the Gleason score, and the grade group. Of all features, panel members rated the biopsy level grade group as the least useful. Given that the grade group is directly computed from the Gleason score and all feedback was presented at the same time, it can be seen as redundant information in the feedback.

While the results of this observer experiment are promising, several limitations have to be addressed. First, we cannot entirely exclude that factors outside of the AI feedback influenced the pathologists’ performance, both positively and negatively. While pathologists did not receive any feedback between the two reads, more experience with viewing cases digitally, the viewer, or in Gleason grading itself could have some influence on the results. Though, we believe that the influence of a second-read is small for several reasons: the majority of pathologists predicted that they scored the same, a significant increase was still found when excluding pathologists who indicated more experience, and the performance on the unseen control cases was also high. Furthermore, AI assistance also improved grading on external data, 6 months after the first experiment, which would be unlikely if the measured effect could be contributed to a higher experience level. For future research, the order in which pathologists received AI assistance could be randomized to further exclude these factors.

Secondly, pathologists were not extensively trained or instructed to use the AI system and were free to use the system in any way during grading. All cases that were graded were also included in the analysis, and we did not allow for a training phase. Pathologists can benefit from an understanding of the global properties of an AI system when such a system is introduced in their grading process; this includes the system’s limitations, its tendency to over- or under grade, and the overall design goal. A training phase at the start of the observer experiment could have increased the use, understanding, and effectiveness of the AI feedback during the grading process and might have led to further increased performance by using the AI system.

Third, in this study, we focused on the assessment of individual biopsies whereas in clinical practice pathologists will examine multiple biopsies per patient. The dataset used to develop the AI system only included one glass slide per patient and the pathologists who set the reference standard evaluated each biopsy individually. An important avenue for future research would be to investigate AI assistance on a patient-level, which allows for new approaches such as automatically prioritizing slides.

Last, the selection of cases under review can influence overall results. We performed our main experiment using a limited set of 100 test and 60 control cases from a single
center in a research setting. The results on the external validation set showed that AI assistance still improved grading, even on data with a different case distribution and reference standard. Nonetheless, for clinical validation of AI tools and their benefit to day-to-day practice, more cases, collected under different settings and from additional centers, should be included.

To the best of our knowledge, our study is the first to show the benefit of AI support for Gleason grading. Ultimately, additional research should determine whether the added benefit of AI assistance results in a stronger prognostic marker for individual patients.

5.5 Acknowledgements

This study was financed by a grant from the Dutch Cancer Society (KWF), grant number KUN 2015–7970. The authors would like to thank Steven Teerenstra for contributing to the statistical analysis, and Kimmo Kartasalo and Henrik Olsson for assisting with the Imagebase dataset.

ISUP Pathology Imagebase Expert Panel
General discussion
6.1 Overview

In this thesis, we set out to improve prostate cancer diagnostics by applying artificial intelligence (AI). With the goal to bring these innovations to the clinic, we cannot discuss AI systems independently of the target domain. Namely, the autonomy of an AI system, its design choices, and its impact on the physician’s workflow vary according to its intended use (Figure 6.1). For example, a system could take a prioritization role and filter biopsies before the pathologist’s review. A large portion of evaluated prostate biopsies does not contain cancer and pathologists’ workload can be drastically reduced if an AI system filters these out. Filtering on its own does not improve the quality of the diagnosis, though. Additionally, the high autonomy of prioritization systems, especially if it screens biopsies, will require rigorous testing to prevent missing clinically relevant cancers.

Mirrored to the screening approach is an AI system that verifies the pathologists’ diagnosis. Unfortunately, this method results in almost zero time gains as the pathologist still performs a regular review and could even reduce efficiency, as more cases will get a second read. Nevertheless, it is a promising approach to increase overall sensitivity, though one should investigate whether this would result in overtreatment of non-significant cancers.

Besides screening and verifying, an AI system can also actively support pathologists during the diagnostic process. This approach is a middle ground in terms of autonomy and influence: The AI system is actively involved, but a pathologist can always override the predictions. By combining the AI system’s and pathologist’s expertise, there is a high potential for improving the quality of the diagnosis. Time-wise, some improvements can be expected as the AI can pre-fill reports and guide pathologists, reducing time spent searching for relevant tissue regions. Overall this AI assistance can be seen as an interactive second pair of eyes: a digital fellow.

The investigation and design of such a ‘digital colleague’ was the main focus of this thesis. Specifically, we aimed to support pathologists using deep learning-based methods when grading prostate biopsies following the Gleason grading system. In this final chapter, we evaluate our work, assess the impact made on the field, and outline future directions.

6.2 Fine-grained segmentations

At the start of this research, automated histological assessment of prostate cancer was still in its infancy – with still a significant focus on feature engineering but the field was transitioning quickly to deep learning-based methods. While promising as
6.2 Fine-grained segmentations

**Prioritize**
- Prioritize biopsies and filter out benign cases.
- Flag malignant cases for review.

**Assist**
- Assist pathologists by highlighting tumor regions and give grade predictions.

**Verify**
- Pathologist diagnoses first. Verify diagnosis with AI output and flag discrepancies.

<table>
<thead>
<tr>
<th>Autonomy and influence of AI</th>
<th>Time savings / efficiency</th>
<th>Improving diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ + +</td>
<td>+ + +</td>
<td>+</td>
</tr>
<tr>
<td>+ + +</td>
<td>+ +</td>
<td>+</td>
</tr>
<tr>
<td>+ + +</td>
<td>+ + +</td>
<td>+</td>
</tr>
</tbody>
</table>

**Figure 6.1:** How an AI system is implemented in clinical practice dictates its autonomy, the possible time savings, and the impact on the diagnosis. There is a trade-off between the potential benefits of such systems versus risk, with more autonomous systems having higher efficiency gains and requiring more stringent validation.

A technique, these existing methods were often very coarse in their output, limiting their application. This coarseness was primarily present in these model’s inability to distinguish between epithelial and stromal tissue, often caused by a lack of detailed training data. Algorithms that could segment with higher precision were desired for correct quantification of tumor areas or for displaying results to pathologists.

Acquiring large amounts of highly detailed data was, and still is, a challenge within computational pathology. Unique to pathology is the ability to use different staining techniques to highlight certain cell types or structures in the tissue. Chapter 2 of this thesis aimed to overcome the annotation burden by employing immunohistochemistry (IHC) to annotate data for model training semi-automatically. Using CK8/18 and P63 to highlight the epithelial tissue and basal cell layer, we could accurately distinguish between the epithelial and stromal tissue. Restaining and registering slides allowed us to map the semi-automatic annotations made using the IHC slides to the H&E versions. This technique showed to speed up data labeling significantly and removed the subjectiveness associated with human-made annotations. Due to this IHC-backed reference standard, the final deep learning model was able to segment the epithelial...
tissue with higher precision than reasonably achievable with just human annotations. The dataset was released publicly for research use as the PESO dataset.\textsuperscript{79}

On its own, a segmentation algorithm for epithelial tissue is of little clinical use. However, the ability to accurately segment epithelial tissue can drastically reduce the annotation burden for higher-level annotation tasks, e.g., cancerous glands. Additionally, the techniques we showed in this chapter are not limited to epithelium segmentation. On the contrary, using IHC to segment tissue semi-automatically is an effective strategy to create detailed annotations and applicable to many tasks within computation pathology.

### 6.3 AI-powered Gleason grading

For patients with prostate cancer, the Gleason score and grade group of a biopsy are among the most important prognostic markers.\textsuperscript{80,96,127} Unfortunately, the grading system suffers from substantial inter- and intraobserver variability,\textsuperscript{20,21,97} and while specialized uropathologists show a higher agreement,\textsuperscript{30} they are not available to every patient. At this point, most existing studies were done on preselected, smaller subimages or tissue microarrays.\textsuperscript{68,69,85,87,128–130} Studies that focussed on full whole slide images either used a small dataset with a subset of Gleason patterns\textsuperscript{88} or prostatectomies.\textsuperscript{27} Chapter 3 investigated whether a deep learning system could grade prostate biopsies at expert levels to make high-quality reproducible Gleason grading more accessible to patients. A substantial part of this project consisted of compiling a new cohort of 5759 prostate biopsies as there was no pre-existing annotated public dataset for Gleason grading available. Annotating this new dataset would quickly require hundreds of pathologist hours. To remedy this, we used a novel pipeline to semi-automatically annotate the data on a gland-by-gland level using the developed epithelium segmentation system (Chapter 2) and an existing tumor detection system.\textsuperscript{38}

The subjective nature of Gleason grading also introduced challenges when assessing the algorithm’s performance. The gold standard for reference standards consists of asking multiple specialized uropathologists to grade biopsies and combining their opinions. Still, while using a panel improves the reference standard’s quality, it is not watertight and remains dependent on a correct assessment by pathologists. It also limits the upper bound of a system’s performance to the experts who set the reference standard. When algorithms improve in performance, we will eventually reach a point where we can argue who is right.

To determine the added value of automated grading for clinical practice, we assessed whether the new technique is non-inferior – and ideally better – compared to the standard practice of care. Thus, we compared the AI system with the grading performance
of a panel of pathologists and residents. We found that, on average, the AI system showed a higher agreement with the experts than the panel, operating at the level of experienced pathologists, and giving concrete proof of the power of this technology. Several other studies have confirmed our findings and demonstrated the high performance of deep learning-based Gleason grading of biopsies. Most notably are the works by Ström et al., published simultaneously in the same journal, and the later work by Nagpal et al. In output and approach, these studies were quite similar to ours, both used deep learning-based models albeit with different architectures. The main differences were twofold: We used an annotation-less approach of collecting the data based on the diagnosis of the original pathologist report and without extensive annotations by pathologists. Secondly, we determined the specimen’s grade group in the same way as pathologists would do in clinical practice, instead of using an additional machine learning model. Thus, as our model detects Gleason patterns on a glandular level, we can directly calculate the primary and secondary patterns and determine the grade group. This technique ensures that the model output is easily interpretable and there is no black box that determines the final diagnosis.

6.4 The PANDA Challenge

Chapter 3 showed the promise for accurate AI-based diagnosis and grading of prostate cancer in biopsies. However, despite our results and that of similar studies, there had been limited proof of generalization to diverse and multinational patient cohorts and reference standards. Additionally, these studies – as common in the field – were developed in a siloed matter: the same researchers who developed the algorithms also validated them. Thus, to assess the maturity of AI for Gleason grading, we required a more independent validation of these new techniques.

Medical imaging competitions, or challenges, have shown great accelerators for AI innovations, with the CAMELYON challenges being a supreme example. However, many of these competitions were also limited in their independent validation and reproducibility of algorithms, raising doubt whether competitions solve the underlying medical problem.

We aimed to address both issues in Chapter 4 through organizing the PANDA challenge: an international competition on AI-based Gleason grading. With a development set of 10,616 biopsies, 2,009 biopsies for internal and external validation, and 1,290 participants, it was the largest competition organized for pathology to date. We required that all challenge participants submitted working algorithms, which were validated entirely independently of the algorithm developers, reducing sources of potential bias and verifying the reproducibility of the proposed solutions. While this
approach requires significantly more compute resources to organize challenges, it is a logical evolution from previous challenge designs and should become the defacto standard in the field. Open research platforms such as grand-challenge.org could play a significant role in this evolvement of challenge designs by making these study setups more accessible.

As all algorithms from PANDA were reproducible, it allowed us to apply algorithms to external cohorts and to assess AI's performance for Gleason grading as a 'whole' instead of individual algorithms. Evaluating a group of algorithms is distinctly different from evaluating just the 'best' algorithm of a competition, which can suffer from limitations and biases. We found that algorithm performance was retained even on cross-continenal datasets, showing that algorithms trained on EU data can generalize to US-based cohorts. Algorithms were highly sensitive for cancer, missing fewer cancers than pathologists, but suffered in terms of specificity on the external validation sets.

Organizing a competition in this way, ensuring independent validation and reproducibility, adds significant evidence of the maturity of AI for automated Gleason grading. In addition, it is one of the most extensive evaluations that can be done within the limits of a retrospective setup. The next logical step is to validate these methods in clinical practice, e.g., by allowing pathologists to use these technologies and through prospective clinical trials.

### 6.5 The added value of AI-assistance

AI successes for medical applications co-occur with a lot of hype and fear that AI could make physicians obsolete. While the diagnostic power of these AI systems is not generally contested, there is still a gap between research results and clinical practice. Even though a system can achieve a high agreement with experts or accurately outline a tumor region, this does not directly translate to improvements in the diagnostic process or better care.

In the research community, there is disproportionate attention for achieving state-of-the-art (SOTA) results: ample new methods are being introduced that claim to achieve SOTA on benchmark datasets. As such results make for good headlines, the reasoning is clear, and of course, we want to drive the field forward to achieve optimum performance. However, it is sometimes questionable whether these slight increases in performance actually result from better algorithm design; It is one reason we opted for a group-based analysis for the PANDA challenge instead of focusing solely on the challenge winner.
To truly evaluate algorithms’ added benefits, we need to assess them in a suitable environment and investigate how we could implement such systems. These evaluations go beyond quality metrics and achieving top scores on benchmarks and should include efficiency assessments and human-computer interaction as well.\textsuperscript{125}

When we completed our work on automated grading (Chapter 3), there were countless other papers on AI for pathology. However, studies showing how these techniques can be used effectively were still rare. There was limited evidence of the merits of AI assistance for the histological assessment of prostate cancer. Some studies have been performed to show the value of AI assistance in detecting prostate cancer,\textsuperscript{120} but no such study existed for Gleason grading. In Chapter 5, we showed that pathologists could actually benefit from AI assistance when diagnosing prostate cancer.

In Chapter 3, the AI system outperformed most pathologists, but the AI-supported pathologists now outperformed both the standalone AI as the pathologists without AI. This manifestation of \textit{pathologist-AI synergy} can be easily explained as both the pathologist and AI system have different strengths and weaknesses. AI systems are consistent, reproducible, and cannot get tired. As a result, all tissue is analyzed consistently, and there is a lower chance that small tumors are missed. Additionally, AI systems offer high precision in quantification, allowing pathologists to better assess the tumor’s extent. On the other hand, algorithms can notoriously fail in the presence of data patterns that were not in the training dataset; examples include artifacts, foreign tissue, or extensive shifts in the stain characteristics. Pathologists will more easily adjust to changes in the staining characteristics and disregard artifacts.

To the best of our knowledge, we were the first to show that AI-assisted can help pathologists perform better at Gleason grading. Shortly after, Steiner et al. also demonstrated the benefits of AI assistance for Gleason grading on a different cohort.\textsuperscript{101} These results are fundamental for the field; we not only need to show the power of AI innovations but more so proof that they can be effectively used in practice. Following this, as was already said for radiologists,\textsuperscript{135} one could say that AI will not directly replace pathologists, but pathologists using AI could replace those who do not.

\section*{6.6 On the importance of data}

An alternative fitting title of this work could have been \textit{“improving prostate cancer grading using big data.”} While technological advances in AI, and specifically deep learning, have allowed researchers to achieve state-of-the-art results, the access to large datasets continues to fuel this revolution. In contrast to the old way of explicit feature engineering, current AI applications for medical imaging regularly use existing architectures or a combination of existing building blocks. The availability of
large datasets, combined with the tools and computing power to harness those, forms the basis of many recent advancements. Additionally, there is a positive correlation between dataset size and model performance. The chapters of this thesis show a similar trend, from 102 cases (Chapter 2) to a cohort of more than 11,000 biopsies (Chapter 4), and can also be seen in many other studies which consistently include large datasets.

With data at the core of AI development, the collection, curation, and intelligent processing of those datasets will only become more and more critical. The results presented in this thesis would not have been possible without novel ways of collecting and transforming the data for model training (Chapter 2 and 3). Even though there is much attention for new methodologies, for many of the current pathology tasks, AI development is probably best started with a good grasp of the data, with the actual deep learning architecture or method following second. Unfortunately, while there is typically fast dissemination of new methodological ideas, the same does not hold for data. Datasets, especially in the medical domain, are often well guarded by the institutions that collected them. Most of this is understandable, either from privacy perspectives or because of the inherent value data holds for research groups. However, because data is so vital, it can also limit some of the high-impact applied research to those who can afford it. By publishing two unique datasets for computational pathology, the PESO and PANDA datasets, we aimed to contribute to the availability of research data and fuel further innovations. Despite that these datasets are new, they have already been downloaded thousands of times. The PESO dataset was used in several internal and external publications, and we expect considerable interest in the PANDA dataset due to its size and application.

### 6.7 Between now and clinical practice

The majority of this thesis has focussed on quantitative analysis of performance, either that of automated systems or AI-assisted pathologists. However, the actual implementation of these techniques goes beyond achieving high performance on independent validation cohorts. With higher autonomy and decision-power of AI systems, we also need to focus on the implementation and robustness of such systems, especially from an end-user perspective.

To start, most of the challenges introduced in Chapter 1 regarding scanner and stain variations remain highly relevant. An algorithm can work flawlessly on a test cohort that originated from the same center as the training cohort, but for real-world scenario’s we need to assure quality among a wide range of image characteristics. Applying techniques such as data augmentation to increase the diversity of the train-
6.7 Between now and clinical practice

ing data and (deep learning-based) normalization can help algorithms become more robust. However, these techniques also have their limits, and building multi-center multi-scanner datasets will remain essential in tackling these problems.

With measures to increase robustness in place, it is still not guaranteed that algorithms will be able to handle all input data. Therefore, it would be sensible to require that quality-assurance (QA) algorithms precede clinical algorithms. Not only from a safety aspect but also to not waste pathologists’ time with faulty predictions due to the garbage-in-garbage-out principle. Such algorithms could check for artifacts, ensure that the tissue is the right type for the clinical algorithm, determine if the stain is of sufficient quality, no tissue is out of focus, and outline valid regions where the clinical algorithm should be applied. Utilizing these techniques, combined with on-site validations, will be vital to retaining model performance across clinics and ensure stable day-to-day operations.

Beyond increasing resilience to data variants and artifacts, we need actual fail-safes that prevent incidents from going unnoticed. Unlike their human colleagues, AI systems in their default forms are often limited in expressing uncertainty and often suffer from catastrophic failure when confronted with data outside the training distribution. Commonly, samples that are not within the set of known classes are still classified as one of those and with high confidence values. When AI is applied outside of the strictly controlled research settings, we will need tools to express certainty, detect these outliers, and flag cases accordingly. Even if a QA algorithm accepted a case for further review, the actual medical device should still alert the pathologist if a diagnosis cannot be made with certainty. The pathologist that uses the AI can then determine not to take the AI’s predictions into account. This out-of-distribution detection and uncertainty estimation is an active field of research and a fundamental ingredient for future AI implementations.

Besides the technical challenges, we should not underestimate the human component when applying AI in clinical practice. In Chapter 5, pathologists used the developed Gleason grading system without any prior training. With proper training and a better understanding of the strengths and weaknesses of an AI system, pathologists will be better equipped to utilize AI in clinical practice. On the other side, we can also adjust how algorithms display predictions to better match pathologists. For example, we can show results in numeric formats, overlays on the tissue, or annotations and outlined regions. The explainability of an algorithm is critical for a clinician’s ability to put the predictions in context and visualizations can benefit or harm this. Unfortunately, in research, these more qualitative results on usability and algorithm output are often hidden in the supplemental or not even focussed on at all. Future research in human-computer interaction, or maybe better described as human-algorithm inter-
action, should guide us in designing systems to match the pathologists that will use them.

To summarize, AI implementation in the pathologist’s workflow co-occurs with many challenges, both technical and from a clinician’s perspective, and not to mention societal issues and regulatory requirements. For Gleason grading specifically, the results of this thesis and related work\textsuperscript{27,86,98,100,101} show that the technical foundation is present, but we also have not yet reached clinical practice. Thus, it is time for prospective validation of these innovations, where we investigate what impact AI can have on clinical diagnostics in real-life scenarios. In such a prospective setting, we can then compare AI-based grading to routine diagnostics, investigate its impact on efficiency, and eventually determine the added benefit for patients. In this “last mile” gap between AI research and clinical implementation,\textsuperscript{144} our challenges will predominantly be focused on everything but the actual algorithm itself.

### 6.8 Beyond human reference standards

This thesis has focused primarily on improving and optimizing the current histopathological diagnostics for prostate cancer. The Gleason grading standard has been around for years, and by adding AI assistance, we aim to improve the diagnostic process through reproducibility, higher quality, and increased efficiency. While these are giant steps and could significantly benefit patients, a cynic could also argue that these are mere ‘process optimizations.’ So what does the future hold for (prostate cancer) diagnostics?

The Gleason system was built around the idea that specific tissue patterns correspond to distinct patient survival stratification.\textsuperscript{155} Humans are good at pattern recognition in a general sense, but it is out of the question that AI can do this at an unprecedented scale and detail. Therefore, the next step forward would be extending or further refining the grading system. Automated quantifications could make grading more reproducible and precise,\textsuperscript{27,100,102} allowing for more fine-grained updates to the Gleason grading system. In addition, utilizing the power of deep learning models to recognize patterns could give further insight into the many variants and subtleties of prostate cancer.\textsuperscript{156}

Detected patterns are not always practical, though; an AI system can also react to specific image characteristics or implicit biases in the data. E.g., suppose images with high-grade cancer originate from a single institution. In that case, the AI system could learn to associate that institution with high-grade tumors instead of using actual tissue patterns for the classification. Evaluation of such systems is therefore not trivial. When we assess AI systems’ performances on tasks using human-based refer-
Figure 6.2: Current iterations of AI for prostate cancer mostly follow the Gleason grading standard to give pathologists feedback (top). New studies focus on skipping the Gleason grading standard and directly predicting prognosis from digitized specimens. Still, these systems are limited to a single modality. Future AI systems could integrate information from multiple modalities, also outside of pathology, to determine the best personalized treatment for a patient (bottom).
ence standards, we are ultimately limited by humans’ ability for pattern recognition. Given the high observer variability of Gleason grading, we will reach this limitation fairly quickly.

A transition has been started in the field that moves away from human-defined subjective reference standards (e.g., a Gleason score) to patient-level ground truths.\textsuperscript{109} For example, using patient survival data as the ground truth could remove the subjectiveness introduced by, e.g., the Gleason score (Figure 6.2). Techniques such as *neural image compression*,\textsuperscript{157} *streaming CNNs*\textsuperscript{158,159} and *multiple instance learning*\textsuperscript{86,160} allow for model training using solely slide or patient-level labels. As part of the PANDA challenge, many teams already demonstrated that training directly using slide-level labels was an effective strategy, albeit pathologists still supplied these labels.

In the future, we will have algorithms that, based on the raw image data, make direct predictions on patient prognosis without any human-defined intermediate. Such digital biomarkers could be the second wave of the AI-powered revolution of clinical diagnostics in pathology. However, defining studies to develop such systems is non-trivial, primarily when based on retrospectively collected data.\textsuperscript{161,162} If we train AI systems on long-term retrospectively collected survival data, the treatment choices based on the pathologist’s original diagnosis could introduce bias and reduce the power of such systems in a prospective setting.

Besides the risk of bias, the second concern of high-level end-to-end systems would be the lack of explainability of such systems. As a result, there is a risk that AI systems could become a black box, even more than some would suggest it is now. A system that detects Gleason patterns and uses this to determine a diagnosis is more investigable than solely a patient-level risk assessment. However, if the black box is validated, tested for bias, achieves high accuracy, and has fail-safes in place, it would benefit the patient to embrace such technology. Adding explainability could therefore be a reverse process. Alternatively to the approach of the Gleason grading system, we will (slowly) move away from manually finding patterns and linking these to prognosis. Instead, an algorithm can directly predict prognosis, and we can try to reverse engineer which patterns contributed to this prediction to allow for human investigation and explainability.\textsuperscript{163–165}

Looking beyond the current generation of AI systems and those that predict slide-level labels is a generation of multimodal AI systems (Figure 6.2). Treatment planning for prostate cancer patients, e.g., by a urologist, is not based solely on the Gleason score. Instead, when determining an individual patient’s optimal treatment, a physician will take all available information into account, including data from pathology, radiology, etc. The next generation of AI systems could aggregate all this information, breaking barriers between modalities and hospital departments to determine an individual
6.9 Conclusion

This thesis described the development and validation of AI-assisted Gleason grading in prostate biopsies. Our work started from the most basic level, finding epithelial tissue in prostate specimens (Chapter 2). We then showed that an AI algorithm achieved pathologist-level performance in a large-scale diagnostic study (Chapter 3). Next, by organizing the most extensive medical image challenge for pathology to date, the PANDA Challenge, we demonstrated that AI for Gleason grading is a mature and promising technology that warrants further validation in prospective settings (Chapter 4). Finally, and perhaps most importantly, we were able to show that AI can indeed improve Gleason grading of prostate biopsies by pathologists (Chapter 5).

The future will hold at least two important directions. First, existing methods can be extended to increase clinical acceptance and streamline implementation, including interpretability, robustness, and interaction with pathologists. These are primarily iterative improvements and not scientific breakthroughs. Still, they are crucial if we want to design AI that actually reaches clinical practice and benefits patients. Second, new AI technologies will allow for novel biomarkers beyond human reference standards and what we use in diagnostics today. Thus, the future is bright for AI-based diagnostics for prostate cancer in pathology, not only from the research and technical perspective but also for pathologists who will harness these new technologies and improve the care they deliver for patients.
Summary
For patients with prostate cancer, the Gleason score and grade group are crucial parts of their diagnosis and significantly influence treatment planning. Pathologists determine the score and group through microscopical assessment of tissue removed from the prostate after, e.g., a biopsy procedure. While the Gleason grading system is the strongest tissue-based marker, it is inherently subjective. Consequently, it exhibits substantial variability between pathologists and even within the same pathologist at different points in time. Prostate cancer diagnostics could thus benefit from more robust, reproducible Gleason grading. Artificial intelligence holds promise for improving this process with its ability to learn from large amounts of data.

Chapter 1 gave a global overview of the fields at the basis of this thesis. We started by explaining the epidemiology of prostate cancer and the diagnostic pathway. Second, we discussed the Gleason grading system and the difficulties that co-exist with this grading standard. Then we explained the digital revolution taking place in pathology and the subsequent introduction of computational pathology. Finally, we set out the overarching aim of this thesis: Improving prostate cancer diagnostics by developing artificial intelligence algorithms that can support pathologists.

Acquiring substantial amounts of highly detailed data for algorithm development is a challenge within computational pathology. Chapter 2 of this thesis aimed to overcome the annotation burden by employing immunohistochemistry (IHC) to annotate data for model training semi-automatically. Using CK8/18 and P63 to highlight the epithelial tissue and basal cell layer, we trained an IHC-based segmentation algorithm. We could then transfer this model’s output using image registration to H&E stained slides and train a highly accurate epithelium segmentation system in H&E. This proved to be an effective method to annotate large slides and drastically reduced the annotation burden for higher-level annotation tasks, e.g., cancerous glands.

Our next challenge was to build a system that could diagnose and grade prostate biopsies at the level of a pathologist (Chapter 3). For this, we required a large dataset of annotated biopsies, which was not readily available. Utilizing the algorithm from Chapter 2, we could semi-automatically annotate a large dataset of almost 6000 biopsies and used this for model training. We compared the AI system with the grading performance of a panel of pathologists and residents. We found that, on average, the AI system showed a higher agreement with experts than the panel, operating at the level of experienced pathologists, and giving concrete proof of the power of AI for Gleason grading.
Additional research was required to further assess the potential of AI for Gleason grading and see if results would translate to more extensive and diverse datasets. In Chapter 4, we introduced the PANDA challenge: an international competition to crowd-source AI innovations for automated Gleason grading. In this challenge, over 1200 algorithm developers worked towards building the best-performing algorithms. Based on performance during the challenge, we selected fifteen algorithms to undergo rigorous independent validation. Unique to this challenge was that we performed the validation entirely blinded to the algorithm developers, minimizing potential bias. We showed that these AI algorithms are capable of pathologist-level performance across cross-continental datasets with various reference standards.

This thesis started with the aim of empowering pathologists with AI to improve the diagnostic process. To this extent, we evaluated the potential benefits of AI when such algorithms are placed in the hands of pathologists. In Chapter 5 we invited pathologists to use our AI algorithm and compared their performance to an unassisted setting. We found that when pathologists used the algorithm, they significantly improved in determining the Gleason grade group. In Chapter 3, the AI system outperformed most pathologists. However, with AI support, pathologists now outperformed both the standalone AI as the pathologists without AI. These results show the existence of a pathologist-AI synergy: algorithms and pathologists both have their strengths and weaknesses that can complement one another.

In Chapter 6 we summarized the main findings of this thesis and outlined several key characteristics of computational pathology that shape current and future projects. First, we highlighted the importance of the availability of large high-quality datasets. Data is at the core of many AI innovations, and the projects conducted as part of this thesis are no exception. Second, we discussed steps that need to be taken to bring these AI innovations to the clinic. Developing robust models that generalize to external labs, introducing fail-safes, and human-algorithm-interaction design are crucial prerequisites for implementing AI in clinical practice. Third and finally, we discussed an outlook for the future, focusing on applications beyond human-based reference standards.

Concluding, we investigated how artificial intelligence can improve prostate cancer diagnostics. Specifically, we showed that AI algorithms are able to achieve pathologist-level performance at a highly complex task such as Gleason grading. AI can assist pathologists, and together they can improve the diagnostic process and ultimately deliver a more accurate diagnosis for patients.
Samenvatting
De Gleason score en graad-groep zijn cruciaal voor patiënten met prostaatkanker; samen met andere klinische parameters sturen ze voor een belangrijk deel het behandelplan. Pathologen bepalen de score en groep op basis van microscopische analyse van weefsel dat is weggenomen uit de prostaat, bijvoorbeeld door middel van een biopsie. Ondanks de cruciale rol van het graderingsysteem is het een subjectieve taak en bestaat er veel variabiliteit tussen pathologen. Deze variabiliteit verlaagt de effectiviteit van het graderingsysteem voor individuele patiënten. Dit proefschrift is het resultaat van ons onderzoek naar hoe kunstmatige intelligentie (AI) pathologen kan ondersteunen in dit proces om zodoende het diagnostisch proces te verbeteren.

In hoofdstuk 1 introduceerden we de verschillende thema’s die aan de basis lagen van dit proefschrift. We gaven een overzicht van de epidemiologie, het diagnostisch proces, het graderingsysteem en de problemen die hierin spelen. Als laatste introduceerden we een recente revolutionaire ontwikkeling binnen de pathologie: de digitalisatie en de uitfasering van analoge microscopen. Er ontstond daarmee ook een nieuw veld waar dit proefschrift een voorbeeld van is: de computatiele pathologie.

Het bepalen van de Gleason score wordt door pathologen uitgevoerd op weefsel dat bewerkt is met de standaard hematoxyline-eosinekleuring (H&E). De ontwikkeling van AI voor dit proces vereist grote hoeveelheden data. Echter, het verzamelen van kwalitatief goede gelabelde data is niet gemakkelijk en vergt doorgaans vele kostbare pathologen-uren. Naast H&E, is het binnen de pathologie ook mogelijk om specifieke weefseltypen aan te kleuren met behulp van immuunhistochemie (IHC). In hoofdstuk 2 gebruikten wij IHC om semiautomatisch data te labelen. Door gebruik te maken van twee kleuringen, CK8/18 en P63, trainden we een systeem dat epitheel kon herkennen in IHC-gekleurde beelden. Door de IHC-beelden uit te lijnen met de H&E versie, waren we in staat een systeem te trainen op H&E. Het trainen vereiste geen handmatige annotaties doordat het systeem kon leren van de annotaties op de IHC-beelden. Deze methode bleek een snelle en effectieve manier te zijn om grote beelden te analyseren, bijvoorbeeld door het automatisch verfijnen van ruwe annotaties gemaakt door mensen.

Onze volgende stap was het ontwikkelen van een algoritme dat volledig automatisch prostaatbiopten kon analyseren, op dezelfde manier als een patholoog zou doen (Houdfstuk 3). Net als de patholoog moest het algoritme op basis van een biopt een Gleason graad-groep afgeven. De benodigde dataset was echter nog niet voorhanden. Met behulp van het algoritme uit hoofdstuk 1 konden we semiautomatisch bijna 6000 biopten van labels voorzien. Het ontwikkelde algoritme bleek te presteren op het ni-
veau van ervaren pathologen en was zelfs beter dan het merendeel van de pathologen in de studie. Het algoritme scoorde zowel hoog in het bepalen van de graad-groep als in het indelen van patiënten in relevante risicogroepen. Deze resultaten lieten zien wat de potentie is van AI voor de diagnostiek van prostaatkanker binnen de pathologie.

Om de resultaten uit hoofdstuk 3 te valideren was er meer onderzoek nodig om te bepalen of deze technologie ook toepasbaar zou zijn in andere centra en settings. In hoofdstuk 4 introduceerden we de PANDA challenge: een internationale programeercompetitie voor AI op het gebied van prostaatkankerdiagnostiek in de pathologie. Aan de competitie deden meer dan 1200 ontwikkelaars mee en de vijftien best presterende algoritmes werden uitgekozen voor uitgebreidere onafhankelijke analyse. Speciaal aan deze opzet was dat de ontwikkelaars geen toegang hadden tot de analyse, waardoor de mogelijkheid tot valspellen minimaal was en de resultaten volledig reproduceerbaar waren. We lieten zien dat AI-algoritmen in staat zijn om op of zelfs boven het niveau van pathologen te presteren, ook wanneer je ze toepast op grote intercontinentale datasets.

De vraag bleef of algoritmen ook daadwerkelijk van nut kunnen zijn voor pathologen en patiënten. In hoofdstuk 5 lieten we pathologen het ontwikkelde algoritme uit hoofdstuk 3 gebruiken. We vergeleken daarna de prestaties van de pathologen met en zonder AI. We vonden dat pathologen significant beter presteerden wanneer ze assistentie kregen van AI. Waar in de originele setting het algoritme nog beter presteerde, bleken de AI-ondersteunde pathologen nu beter te presteren dan de AI op zichzelf én beter dan pathologen zonder AI. Deze resultaten laten zien dat er een synergie kan zijn tussen patholoog en AI: beide hebben hun sterktes en zwaktes en juist de combinatie resulteert in de beste diagnose voor een patiënt.

Hoofdstuk 6 vat de resultaten van dit proefschrift samen en gaf een overzicht van belangrijke thema's die nu en in de toekomst gaan spelen binnen de computatieele pathologie. Als eerste stipten we het belang aan van de beschikbaarheid van grote en kwalitatief goede datasets. Data speelt een cruciale rol binnen de AI en dit proefschrift was daar geen uitzondering op. Als tweede bespraken we stappen die nog genomen moeten worden om deze innovaties naar de klinische praktijk te brengen. Robuuste algoritmen die generaliseren naar externe labs, adequate veiligheidsmaatregelen tegen fouten die kunnen optreden, en een goed design voor de interface tussen algoritme en arts zijn van cruciaal belang voor implementatie en gebruik in de dagelijkse praktijk. Als derde en laatste focusten we op het gebruik van referentiestandaarden op
patiëntniveau. Er is veel potentie voor algoritmes die, op basis van informatie van meerdere disciplines, het ziektebeloop voorspellen voor een individuele patiënt.

In dit proefschrift lieten we zien dat kunstmatige intelligentie de diagnostiek van prostaatkanker kan verbeteren. Algoritmes zijn in staat om op het niveau van ervaren pathologen te presteren, zelfs bij een complexe taak als het Gleason graderingsysteem. AI kan pathologen ondersteunen, en juist deze samenwerking kan de diagnostiek verbeteren met als uiteindelijke doel: een accuratere diagnose voor patiënten.
Publications
Papers in international journals


Publications

Challenge consortium. “Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer: the PANDA challenge”. Accepted for publication in Nature Medicine. 2021

Papers in conference proceedings


Abstracts in conference proceedings


Published datasets

**The PESO dataset**, 102 digitized H&E stained prostatectomy slides with corresponding training masks, available under a Creative Commons BY-SA-NC 4.0 license.


**The PANDA Challenge dataset**, 10,616 digitized H&E stained prostate biopsies with corresponding training masks, available under a Creative Commons BY-SA-NC 4.0 license.


Media

P. Sewuster. “De computer is slimmer dan de patholoog”. In: BNR Nieuwsradio (Jan. 8, 2020)

“Nederlandse onderzoekers gebruiken AI om agressiviteit tumor te bepalen”. In: NU.nl (Jan. 8, 2020)

N. Carne. “Deep learning to aid prostate diagnosis - AI finding a new niche in pathology”. In: Cosmos (Jan. 12, 2020)

“Vliegende auto’s op de CES en Nederlands algoritme herkent prostaatkanker beter dan artsen”. In: NOS op 3 Tech Podcast (Nov. 9, 2020)

“AI-systeem kan agressiviteit prostaatkanker bepalen”. In: ICT & health (Jan. 10, 2020)

N. Korteweg. “De software die slimmer is dan de dokter”. In: NRC (Jan. 31, 2020)

H. Tangerman. “Digitale Dokters - Wat AI voor de medische wereld kan betekenen”. In: KLIK (May 1, 2020)

Awards


RIHS Science Award 2020 for best peer-reviewed scientific publication within RIHS: “Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study.”
Research Data Management
The research project described in this thesis uses an extensive amount of data to train and evaluate several machine learning algorithms. This data consists of three main components: (1) digitized whole-slide images of patient tissue, (2) pixel-level annotations associated with these images, and (3) labels that describe these images at patient and slide level. We adhere to the FAIR data principles (findable, accessible, interoperable, and re-usable) whenever possible.

We strictly follow the Radboud University Medical Center regulations regarding the origin, ownership, and permission to use this data. For each of the datasets used in this research, we have entered data license agreements with the datasets’ stakeholders, if needed, and have obtained permission to use the data for research purposes. The overall study was approved by the institutional review board of Radboud University Medical Center (IRB number 2016–2275).

To protect patients’ privacy rights, all data used within the context of this research project has been subject to pseudo- or anonymization. This process ensures that personally identifiable information is removed or replaced by artificial identifiers, or pseudonyms, before conducting any of the experiments described within this thesis. For the whole-slide images, image data was extracted patch-by-patch and exported to new files to remove any metadata present in the original files. All this data is securely stored within the Radboudumc storage system.

The two primary datasets collected as part of this thesis, specifically the PESO\(^1\) (Chapter 2) and PANDA\(^2\) (Chapter 3 and 4) datasets, were made publicly available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license. These datasets can be downloaded from the respective repositories for research use. Before publication, we removed patient identifying information from the datasets.

Most scientific experiments conducted within the context of this research project using patient data have been executed exclusively within the Radboudumc IT infrastructure. The exception to this rule were experiments related to the PANDA Challenge, which have been conducted in three respective cloud environments (Kaggle through Google Cloud Platform, Google Cloud Platform, and the Puhti compute cluster). All data that was used in these experiments first underwent a process to remove patient identifying information before it was shared with partners.

---

\(^1\)http://doi.org/10.5281/zenodo.1485967

\(^2\)https://panda.grand-challenge.org
PhD Portfolio
**Name:** Wouter Bulten  
**PhD period:** 27-03-2017 until 31-03-2021

<table>
<thead>
<tr>
<th>Courses &amp; Workshops</th>
<th>Year(s)</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction day Radboudumc</td>
<td>2017</td>
<td>0.25</td>
</tr>
<tr>
<td>Graduate School specific introductory course</td>
<td>2017</td>
<td>0.75</td>
</tr>
<tr>
<td>Scientific Writing for PhD Candidates</td>
<td>2017</td>
<td>3.00</td>
</tr>
<tr>
<td>NFBIA Summer School</td>
<td>2017</td>
<td>1.25</td>
</tr>
<tr>
<td>Basic Pathology Introduction Course</td>
<td>2017</td>
<td>0.50</td>
</tr>
<tr>
<td>OostNL Customer Track, entrepreneurship</td>
<td>2019</td>
<td>2.00</td>
</tr>
<tr>
<td>The Art of Presenting Science</td>
<td>2019</td>
<td>1.50</td>
</tr>
<tr>
<td>Scientific Integrity course</td>
<td>2019</td>
<td>1.00</td>
</tr>
<tr>
<td>Radboud Talks 2020</td>
<td>2020</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**Seminars & Lectures**

<table>
<thead>
<tr>
<th>Seminars &amp; Lectures</th>
<th>Year(s)</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radboud Research Rounds</td>
<td>2017-2019</td>
<td>0.25</td>
</tr>
<tr>
<td>DIAG Discussion Hour</td>
<td>2017-2021</td>
<td>5.00</td>
</tr>
<tr>
<td>Computational Pathology Meeting</td>
<td>2017-2021</td>
<td>5.00</td>
</tr>
<tr>
<td>Meeting on Academic Integrity and Authorship</td>
<td>2017</td>
<td>0.20</td>
</tr>
<tr>
<td>Radboud New Frontiers 2018†</td>
<td>2018</td>
<td>0.50</td>
</tr>
<tr>
<td>Health Valley Event</td>
<td>2019</td>
<td>0.25</td>
</tr>
<tr>
<td>NWO Wetenschapper 2030: Evolutie of Revolutie</td>
<td>2019</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**Symposia & Conferences**

<table>
<thead>
<tr>
<th>Symposia &amp; Conferences</th>
<th>Year(s)</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIHS PhD Retreat*</td>
<td>2017-2019</td>
<td>2.25</td>
</tr>
<tr>
<td>Computational Pathology Symposium 2017</td>
<td>2017</td>
<td>0.25</td>
</tr>
<tr>
<td>SPIE Medical Imaging 2018†</td>
<td>2018</td>
<td>2.00</td>
</tr>
<tr>
<td>Medical Imaging with Deep Learning (MIDL) 2018†</td>
<td>2018</td>
<td>1.25</td>
</tr>
<tr>
<td>USCAP 108th Annual Meeting*</td>
<td>2019</td>
<td>2.00</td>
</tr>
<tr>
<td>Medical Imaging with Deep Learning (MIDL) 2019†</td>
<td>2019</td>
<td>1.25</td>
</tr>
<tr>
<td>PANDA Challenge workshop (organizer) at MICCAI 2020*</td>
<td>2020</td>
<td>2.00</td>
</tr>
<tr>
<td>AICI Forum 2020†</td>
<td>2020</td>
<td>0.25</td>
</tr>
<tr>
<td>DELAB-Fachtagung für Ärztinnen und Ärzte‡</td>
<td>2020</td>
<td>0.25</td>
</tr>
<tr>
<td>RDNL Together we share</td>
<td>2020</td>
<td>0.25</td>
</tr>
<tr>
<td>AI &amp; IA - Imaging Architecture in het tijdperk van AI</td>
<td>2020</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**Other**

<table>
<thead>
<tr>
<th>Other</th>
<th>Year(s)</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Learning Journal Club</td>
<td>2017-2021</td>
<td>4.00</td>
</tr>
<tr>
<td>Reviewing scientific publications</td>
<td>2017-2021</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Teaching & Supervision**

<table>
<thead>
<tr>
<th>Teaching &amp; Supervision</th>
<th>Year(s)</th>
<th>ECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching assistant at Intelligent Systems in Medical Imaging</td>
<td>2017-2018</td>
<td>1.00</td>
</tr>
<tr>
<td>Supervision of master students</td>
<td>2018-2021</td>
<td>4.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>43.95</td>
</tr>
</tbody>
</table>

* Poster/laptop presentation,*  
* Oral presentation,*  
‡ Invited talk.
Bibliography
Bibliography


Bibliography


[175] “Vliegende auto’s op de CES en Nederlands algoritme herkent prostaatkanker beter dan artsen”. In: NOS op 3 Tech Podcast (Nov. 9, 2020).


Acknowledgements
Een promotietraject kan soms een individuele bezigheid lijken: er staat één naam op het proefschrift, artikelen hebben (doorgaans) één eerste auteur en het geheel moet als bewijs dienen dat je als onafhankelijke individuele onderzoeker kan opereren. In de praktijk is doorgaans het omgekeerde waar en is wetenschappelijke output een resultaat van een team. Dit proefschrift was er niet zonder hen, van iedereen die dicht betrokken was tot iedereen die zijdelings een (kleine) bijdrage heeft geleverd.

Ik wil graag beginnen met mijn dank uitten naar mijn promotieteam, beginnend met mijn promotoren Jeroen en Bram. Soms van dichtbij maar vaker vanaf een afstandje gaven jullie begeleiding om mij en de rest van het team scherp te houden. Jeroen, jou wil ik nog in het bijzonder bedanken voor de welkome groep die je hebt opgezet. Vanaf het begin van mijn PhD voelde het als een hechte groep waar ook het welbevinden van iedereen belangrijk was. Je hebt betrokkenheid én tegelijkertijd altijd druk zijn tot een kunst verheven.

Aan het begin van mijn onderzoek was mijn pathologiekennis volledig afwezig. Gelukkig had ik een persoonlijke leermeester die mij, op zeer korte termijn, wegwijzig maakte binnen de pathologie. Christina, jouw expertise en enthousiasme heeft mij laten zien hoe mooi het veld van pathologie is. Tegelijkertijd was het ook een goede realisatie qua complexiteit die voor me lag. Ondanks dat onze samenwerking op het einde wat meer op afstand was, heeft die beginperiode een onmisbare solide basis gegeven voor de rest van mijn promotietraject.

Als laatste lid van het promotieteam, maar zeker niet de minste, wil ik Geert bedanken als mijn dagelijkse begeleider. Ik kan in volledige eerlijkheid zeggen dat ik niet een betere begeleider kon wensen. We zeiden het al bij jouw nominatie van supervisor of the year: je hebt een gouden combinatie van expertise in het veld en oog voor het persoonlijke traject van je promovendi. Er waren meerdere, cruciale, momenten waarbij ik bij jou terecht kon voor een stukje bezinning of relativering. Bedankt!

Within DIAG and specifically CPG, there are many people who I have shared many beautiful moments with, in no particular order: Francesco, Witali, Maschenka, Leander, Anton, Tariq, Jasper, Esther, Leslie, Merijn, Miriam, Elke, Babak, Zaneta, Oscar, Henkjan, Maud, Colin, James, Ajay, Erdi, Cristina, Matin, Luuk, Thomas, Kevin, Kicky, Ecem, Kiran, Ioannis, Sjoerd, Coen, Sil, Gabriel, Bart, Alejandro, Clarisa, Jonas, Paul, and many others, thank you for the great time!

Some I want to mention explicitly, starting with my two paranymphs. David, in many of my PhD adventures, you were there as well: from exploring the crazy (under-
ground) world of Houston with Midas, Daan, Ajay, and Thomas, to organizing the best DIAG weekend. During the course of our PhD projects, we were not only office neighbors but we also shared many hotel rooms. Even in my final year, we started a new adventure together. John-Melle, naast alle U-Net discussies waren de wijnproeverijen en onze home automation gesprekken een mooie afwisseling van al het PhD werk. Wanneer is de volgende proeverij? Péter, as the father of our code library, you helped me tremendously in the early phases of my PhD. Meyke, met jouw opmerking op mijn eerste dag “we moeten niet de new guy het idee geven dat we hier altijd de hele avond doorwerken” zette je je meteen neer als iemand die hart heeft voor het groepsgevoel. Hans, als mijn mede-prostaat-promovendus-bij-Geert was jij er altijd voor eerste-hulp-bij-pathologie-vragen of gevraagd én ongevraagd advies over fontjes en design. Daan, onze Houston avonturen waren natuurlijk al legendarisch, maar nog herkenbaarder waren de gesprekken over welk ingrediënt of techniek je nu weer opgepikt had. Thomas, samen met Hans zetten wij de eerste stapjes in de wereld van de value propositions en elevator pitches. Daarnaast was je altijd beschikbaar om te helpen. Mart, als gangmaker en inmiddels webmaster van het webteam waren onze woensdagochtenden met Meyke altijd een lichtpuntje. We gaan het niet eens worden over drankkeuze, maar je kookkunsten maken veel goed. Caner, I will give you a 6.5 for name pronunciation efforts but a 9 for being my office neighbor in the last phase of my PhD. Midas, ik ben het met je eens dat wij samen met Anton, David en Ioannis, het object beste DIAG-weekend ooit organiseerden, met vele mooie press-momenten. Meer op de achtergrond waren Rob, Irene en met name Jeffrey; jullie waren instrumenteel bij de verzameling van data die instrumenteel waren voor mijn hoofdstukken. Hetzelfde geldt voor iedereen die op de afdeling pathologie onderzoek faciliteert, en de studenten die veel van de ‘heavy lifting’ in het annoteren hebben opgepakt: Nikki, Frederike en Milly.

In the last year, most interactions with colleagues moved to the digital domain. As a side effect, this also meant that colleagues worldwide felt almost as close-by as those living in Nijmegen. I’ve had the pleasure to work with brilliant people in the PANDA consortium. Kimmo, Martin, Peter, I’ve learned a lot from you, and it was a great pleasure to work on this large project together (the 18.000 Slack messages show how big of a project it was). Cameron and Kunal, having you in the project team strengthened our project, and I’ve learned a lot from you through our various (sometimes hypothetical) discussions. I also want to thank all the pathologists who contributed to this work and who (remotely) joined the various reader studies. Specifiek wil ik ook nog even Hester en Robert uitlichten die, samen met Christina, de referentiestandaarden hebben bepaald voor meerdere van de datasets van dit proefschrift.
Naast iedereen die ik via mijn onderzoek heb mogen ontmoeten, heb ik om mij heen ook veel gehad aan mijn familie en vrienden. Een paar van hen wil ik nog even extra uitlichten. Om te beginnen met Martijn, jij gaf mij in een periode dat ik het nodig had de gouden tip om naar het project te kijken waar dit proefschrift het resultaat van is. Robert-Jan en Remco, voor jullie is mijn antwoord “het is druk” waarschijnlijk té normaal geworden. Fijn dat jullie er waren voor de nodige afleiding en ruggespraak. Gaby, onze biertjes in de Aesculaaf waren een mooi excuus om te praten over alles wat ons dwars zat tijdens onze promoties. Bart en Theo, ik hoop dat dit proefschrift ook onder ons lijstje van ‘mooie dingen maken’ mag vallen!

Niet te vergeten in dit alles is Huize Houtlaan, inclusief mijn ouders Erik & Anita, en Anne & Claudius. Jullie stonden altijd klaar voor me en leefden intens mee: vanaf het eerste moment toen ik moest kiezen of ik deze stap ging maken tot het in spanning wachten hoe het met mijn artikelen ging. Anita, je bent altijd zo lief, betrokken en geïnteresseerd. Erik, de coaching rol die je op je neemt is voor mij in meerdere situaties van doorslaggevend belang geweest. Als ik ergens hulp bij nodig heb, en het ligt in jullie macht dan doen jullie het. De steun die ik vanuit jullie voel is altijd een bijzondere solide basis geweest, en voor mij van onschatbare waarde.

Als laatste is er één iemand die mijn promotie, na mij, van zo dichtbij heeft beleefd. Enkele dagen voor de start van mijn PhD leerde ik je kennen, en ik zag jou voor het eerst “bij toeval” na mijn tweede officiële dag. Sindsdien ben je bij elk dal en hoogtepunt geweest. Van de slopende deadlinedruk en late avonden, tot aan het vieren van lang gehoopt accept. Het ging zelfs zo ver dat ik een ontbijt in een spa in Costa Rica moest onderbreken omdat ik toch even met Nederland moest bellen. Ik weet dat het niet altijd makkelijk is geweest, zeker met een partner die werkt in hetzelfde veld, maar je stond altijd voor me klaar. We maakten vaak grapjes over co-auteurschappen, maar met alle uren die je hebt besteed, zowel inhoudelijk als daarbuiten, zou Daisy Ermers daar zeker bij hebben kunnen staan.
Wouter Bulten was born in Nijmegen, The Netherlands, in 1991. He studied Artificial Intelligence at Radboud University, for which he obtained his master's degree cum laude in 2015.

His master thesis, “Human SLAM: Simultaneous Localisation and Configuration (SLAC) of indoor Wireless Sensor Networks and their users”, focussed on privacy-aware indoor localization of devices and humans by translating techniques from robotics to the Internet of Things. This project was performed as part of the European ITEA3 Pro-Heal project, in collaboration with Almende and DoBots, Rotterdam, The Netherlands. The results of the thesis were presented at the 1st IEEE International Conference on Internet-of-Things Design and Implementation, Berlin, Germany. For this work, Wouter was awarded the University Study Award in 2016.

Starting from 2011, he worked as a software developer during his studies at 123test, Nijmegen, The Netherlands, where he continued working as a Data Scientist after his graduation. His work focussed primarily on backend software development, architectural/API design, and the development of new systems to improve the administration of psychological tests. Besides this position, he worked as an independent software developer for a wide range of clients.

In 2017, Wouter joined the Computational Pathology Group and the Diagnostic Image Analysis Group under the supervision of Geert Litjens as a Ph.D. candidate on deep learning for improved prognostics in prostate cancer, funded by the Dutch Cancer Foundation (KWF). His research has been published in key journals in the field, and gathered national and international media attention.