4-2-2020 12:00 PM

Machine Learning for Prostate Histopathology Assessment

Wenchao Han, *The University of Western Ontario*

Supervisor: Ward, Aaron D., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Medical Biophysics

Abstract

Pathology reporting on radical prostatectomy (RP) specimens is essential to post-surgery patient care. However, current pathology interpretation of RP sections is typically qualitative and subject to intra- and inter-observer variability, which challenges quantitative and repeatable reporting of lesion grade, size, location, and spread. Therefore, we developed and validated a software platform that can automatically detect and grade cancerous regions on whole slide images (WSIs) of whole-mount RP sections to support quantitative and visual reporting. Our study used hæmatoxylin- and eosin-stained WSIs from 299 whole-mount RP sections from 71 patients, comprising 1.2 million 480μm×480μm regions-of-interest (ROIs) covering benign and cancerous tissues which contain all clinically relevant grade groups. Each cancerous region was annotated and graded by an expert genitourinary pathologist. We used a machine learning approach with 7 different classifiers (3 non-deep learning and 4 deep learning) to classify: 1) each ROI as cancerous vs. non-cancerous, and 2) each cancerous ROI as high- vs. low-grade. Since recent studies found some subtypes beyond Gleason grade to have independent prognostic value, we also used one deep learning method to classify each cancerous ROI from 87 RP sections of 25 patients as each of eight subtypes to support further clinical pathology research on this topic. We cross-validated each system against the expert annotations. To compensate for the staining variability across different WSIs from different patients, we computed the tissue component map (TCM) using our proposed adaptive thresholding algorithm to label nucleus pixels, global thresholding to label lumen pixels, and assigning the rest as stroma/other. Fine-tuning AlexNet with ROIs of the TCM yielded the best results for prostate cancer (PCa) detection and grading, with areas under the receiver operating characteristic curve (AUCs) of 0.98 and 0.93, respectively, followed by

fine-tuned AlexNet with ROIs of the raw image. For subtype grading, fine-tuning AlexNet with ROIs of the raw image yielded AUCs $\geqslant$ 0.7 for seven of eight subtypes. To conclude, deep learning approaches outperformed non-deep learning approaches for PCa detection and grading. The TCMs provided the primary cues for PCa detection and grading. Machine learning can be used for subtype grading beyond the Gleason grading system.

Keywords

Prostate cancer, radical prostatectomy, cancer grading, cancer subtype grading, whole slide image, whole-mount, digital pathology, machine learning, deep learning, nuclei segmentation, texture feature analysis

Summary

Prostate cancer (PCa) is the most prevalent non-skin cancer for Canadian men. Radical prostatectomy (RP) is a surgery that removes the prostate. It is considered to be one of the most effective treatments for PCa patients. However, approximately 30% of patients suffer from recurrence after surgery. Post-surgery patient care, which is advised by pathology reporting on RP specimens, is essential and can be life-saving. Pathology reporting usually provides information such as the presence of tumours, tumour location, and Gleason grade (i.e. a numerical indicator reflecting the aggressiveness of the tumour). However, current pathology interpretation on RP sections is typically qualitative and subject to intra- and inter-observer variability, which challenges quantitative and repeatable reporting of lesion grade, size, location, and spread. Graphical and quantitative reporting, which annotates and grades each tumour with quantitative tumour information associated, can potentially resolve those challenges to better advise post-surgery patient care and pathological studies. However, manually annotating and grading each cancerous region is not feasible in the standard clinical workflow, because tissue sections are enormous under the microscope. Therefore, there is an unmet need for an automatic system that can label and grade cancerous regions on whole slide images (WSIs) of RP specimens. The advancement of scanning technology enables the digitization of WSIs with enough resolution for pathology evaluation. Machine learning is a technique which can identify objects by training the machine with human-labeled examples. Previous research has demonstrated the feasibility of using machine learning to identify and grade regions of interest of prostate tissues. However, detecting and grading each tumour on whole-mount WSIs is still challenging due to the large sizes of high-resolution WSIs, and the staining variability across WSIs. We developed and validated a machine learning based system against expert annotations for PCa detection and grading on 299 whole-mount WSIs,

and for PCa subtype grading on 87 whole-mount WSIs. The systems yielded areas under the receiver operating characteristic curve (AUCs) of 0.98 and 0.92 for PCa detection and grading, respectively, and AUCs $\geqslant$ 0.7 for seven of the eight subtypes. This demonstrates state-of-the-art performance and the potential for clinical translation of this tool.

Co-Authorship Statement

This thesis is presented in an integrated article format. The chapters are based on the following publications that are either under review or in preparation for submission:

Chapter 2: Wenchao Han, Carol Johnson, Mena Gaed, Jose A. Gomez-Lemus, Madeleine Moussa, Joseph Chin, Stephen Pautler, Glenn Bauman, and Aaron Ward, "Automatic cancer detection on digital histopathology images of mid-gland radical prostatectomy specimens," Journal of Medical Imaging (under review).

My contribution to this work included designing the experiments, designing the algorithms for tissue component segmentation, conducting all the experiments, analyzing and interpreting the results, and drafting and submitting the manuscript. C. Johnson helped in the implementation of the support vector machine classifier. M Gaed, J. A. Gomez, and M. Moussa contributed to collecting the histology data, and performing annotation. J. A. Gomez and M. Moussa helped in interpreting the results. J. Chin and S. Pautler recruited subjects and performed radical prostatectomies to provide specimens for analysis. A. Ward contributed to defining the research questions, designing and analyzing the experiments, interpreting the results and drafting the manuscript. All authors helped in reviewing the manuscript. This work was performed under the supervision of A. Ward. G. Bauman was the principal investigator of the grant that funded the Image Guidance for Prostate Cancer trial, from which the histology data was obtained.

Chapter 3: Wenchao Han, Carol Johnson, Mena Gaed, Jose A. Gomez-Lemus, Madeleine Moussa, Joseph Chin, Stephen Pautler, Glenn Bauman, and Aaron Ward, "Histologic tissue

components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens," Scientific Reports (under review).

My contribution to this work included defining the research questions, designing and conducting all the experiments, analyzing and interpreting the results, and drafting and submitting the manuscript. C. Johnson helped in the implementation of the support vector machine classifier. M Gaed, J. A. Gomez, and M. Moussa contributed to collecting the histology data, and performing the annotations. J. A. Gomez and M. Moussa helped in interpreting the results. J. Chin and S. Pautler recruited subjects and performed radical prostatectomies to provide specimens for analysis. A. Ward contributed to defining the research questions, designing and analyzing the experiments, interpreting the results and drafting the manuscript. All authors helped in reviewing the manuscript. This work was performed under the supervision of A. Ward. G. Bauman was the principal investigator of the grant that funded Image Guidance for Prostate Cancer trial, from which the histology data was obtained.

Chapter 4: Wenchao Han, Michelle Downes, Theodorous van der Kwast, Joseph Chin, Stephen Pautler, and Aaron Ward, "Automatic prostate cancer sub-grading on digital histopathology images of radical prostatectomy specimens," Pathology Informatics (in preparation).

My contribution to this work included defining the research questions, designing and conducting all the experiments, analyzing and interpreting the results, and drafting the manuscript. M. Downes and T. van der Kwast performed annotation, and helped in interpreting the results. J. Chin and S. Pautler recruited subjects and performed radical prostatectomies to provide specimens for analysis. A. Ward contributed to designing and

analyzing the experiments, interpreting the results and drafting the manuscript. All authors helped in reviewing the manuscript. This work was performed under the supervision of A. Ward.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Acc.	accuracy
AUC	area under the receiver operating characteristic curve
BPH	benign prostatic hyperplasia
CNN	convolutional neural network
CV	cross-validation
EPE	extraprostatic extension
G	Gleason grade
GS	Gleason score
FisherC	Fisher linear discriminant classifier
FNR	false positive rate
FPR	false negative rate
GLCM	grey-level co-occurrence matrix
GLRLM	grey-level run-length matrix
H&E	hematoxylin and eosin
ISUP	International Society of Urological Pathology
LoglC	logistic linear classifier
LOO	leave-one-out
LOPO	leave-one-patient-out
PCa	prostate cancer
PIN	high-grade prostate intraepithelial neoplasia
PLND	pelvic lymph node dissection
PSA	prostate-specific antigen
PSADT	prostate-specific antigen double time
PSM	positive surgical margin
pT	pathological stage
RBC	red blood cell
RGB	red green blue
ROC	receiver operating characteristic curve
ROI	region-of-interest
RP	radical prostatectomy
SIFT	scale-invariant feature transform
SVI	seminal vesicle invasion
SVM	support vector machine
TMA	tissue microarray
TCM	tissue component maps
WSI	whole slide image

List of Appendices

Chapter 1

1 Introduction

Currently, pathology interpretation of prostate cancer (PCa) in removed specimens from radical prostatectomy (RP) is primarily qualitative, leading to challenges in quantitative and repeatable interpretation of tumour size, location, aggressiveness and spread after surgery. Annotating and grading each tumour on the tissue of RP sections allow for graphical and quantitative pathology reporting, which potentially benefit post-surgery risk management, recurrence prediction, follow-up treatment planning, and pathology related studies.

However, manual annotation of each tumour on the digital histopathology images and grading each of them based on the Gleason grading system [1] (i.e. stratifying the tumours based on their morphological patterns, which reflect the aggressiveness of the tumours) are not currently performed in the standard clinical workflow since they require expert-level knowledge, and are time consuming (e.g. one patient case can take up to 70 working hours for a trained physician to finish [2]). Therefore, there is an unmet need for a system that can automatically delineate and grade tumours on digital histopathology images of RP tissue sections. However, there are key challenges in developing and validating such a computational system: 1) since the tissue is scanned at very high resolution, the image is usually large in size (e.g. 4 – 5 billion pixels per mid-gland whole-mount whole slide image (WSI)), requiring the system to process large amount of image data efficiently; 2) the tissue appearance is largely heterogeneous in both cancerous and non-cancerous tissues, thus the system needs to be able to identify various

patterns for each tissue type for correct classification; and 3) since there are large staining variations across the WSIs within and across patients, the system needs to be robust enough to overcome staining variability to achieve consistent performance for translational applications.

This thesis describes our work on developing and validating machine learning based systems for automatic PCa (1) detection, (2) grading, and (3) subtype grading on whole-mount WSIs of RP sections.

1.1 Background

1.1.1 Prostate cancer epidemiology

PCa has surpassed lung cancer as the most frequently diagnosed non-skin cancer in Canadian men since 1998 [3]. From a recent statistical report, in 2017, there were an estimated 21,300 men diagnosed with prostate cancer, representing 21% of the total new cases in men. It is the third most common cause of death from cancer in men in Canada, which accounts for approximately 10% of all cancer deaths in men. It is estimated that about one in seven Canadian men is expected to be diagnosed with prostate cancer during his lifetime and one in 29 will die from it [4].

Prostate cancer is a highly variable disease, such that a man is more likely to die with prostate cancer than of it. Despite the high incidence of prostate cancer, the five-year overall survival rate is 95% [4]. On the other hand, there are lethal forms of PCa that can be fatal. Early detection and treatment play important roles for PCa patient care. In one study [5], radical treatment (i.e. radical radiotherapy and radical prostatectomy) have

been proven to be associated with reduction of death. In a controlled trial [6], RP was found to reduce the death rate cause by PCa for younger (i.e. age $\leqslant$ 65) PCa patients.

Although RP is considered to be an effective treatment for PCa, there is still a large number of patients suffering from cancer recurrence and metastasis after the surgery [7]. Post-surgery patient care is important and can be life-saving. For example, adjuvant local/systemic therapy may benefit men with extraprostatic extension (EPE) [6].

Pathological assessment of the prostate tissue of RP specimens plays an essential role for patient post-surgery follow-up, which provides information for recurrence prediction, prognosis, and supports selection and guidance of post-surgery treatment. However, currently, pathological interpretation is primarily qualitative, leading to many clinical challenges in reporting pathological parameters (e.g., Gleason score, tumour volume, pathological stage, and tumour zonal location) accurately, efficiently, and consistently. In addition, many prognostic predictors of pathological parameters, such as maximum diameter of the dominant tumour and the presence/absence of intraductal carcinoma etc. have not been fully explored because only a limited number of studies exist. Further study of these questions required a large amount of data (i.e., pathological reporting on those parameters) from large patient cohorts across multiple institutions. These studies are limited primarily due to technical difficulties of accurately reporting those parameters, which usually requires contouring and grading of each tumour on the tissue sections. This is not feasible in current clinical practice.

Thus, the key challenge lies in having graphical pathology reporting with annotations for each tumour at high-precision on RP tissue sections, which yields

associated quantitative pathology reporting, for quantitative and repeatable interpretation of lesion size, location, aggressiveness and spread.

1.1.2 Radical prostatectomy

RP is a surgery that removes the prostate and/or its surrounding tissues through open surgery (e.g. retropubic or perineal surgery) or minimally invasive surgery (e.g. laparoscopic robot-assisted surgery). RP is the most common treatment for patients with organ-confined PCa, which is performed on approximately 40% of PCa patients annually [8] and is considered as gold-standard treatment for clinically localized PCa (i.e. clinical stage $\leqslant$ T2) [9]. Although there are many potential surgical complications such as urinary incontinence, erectile dysfunction, and impotence, there is no evidence showing that any treatment provides better disease control than radical prostatectomy for the primary tumour and distant metastases [10].

1.1.2.1 Indication for radical prostatectomy.

RP is an appropriate treatment for patients who have clinically organ-confined PCa. However, RP should be recommended for patients who have life expectancy of $\geqslant$ 10 years, considering the potential perioperative morbidity [11].

PCa patients of very low to intermediate – risk are candidates for RP, with clinical stage from T1 to T2 (i.e. organ-confined). Other criteria are used for the selection and guidance of RP, such as life expectancy, the predicted probability of lymph node metastasis, prostate-specific antigen (PSA) level, and pathological grade group of biopsies [11].

Some patient groups of high or very high risk may benefit from RP with pelvic lymph node dissection (PLND) especially for the younger and healthier patients. RP with PLND is also an option used as salvage therapy (i.e. therapy for patients having biochemical recurrence after definitive treatment) for patients with post-irradiation recurrence, whose original clinical stage are T1 – 2 with life expectancy of 10 years or more, PSA < 10 ng/ml, and low suspicion of metastases [11].

1.1.2.2 Post-surgery patient care

Although RP is the most common treatment for organ-confined PCa with the advantages of reduction of mortality rate from the disease and the intent of curing the disease, recurrence after the surgery is common. Approximately one third of RP patients have experienced recurrence within 10 years [12]. Also, Pound et al. found that 45%, 77%, and 96% of the patients experienced recurrence within the first 2, 7, and 10 years respectively [13]. In addition to the recurrence, some RP patients may have adverse pathological features, or positive lymph nodes during or after surgery [11], which indicates high risk of biochemical recurrence.

After RP, patients may be recommended for active surveillance, which is a watch and wait strategy with frequent checking for signs of recurrence. Also, adjuvant therapy may be an option, which administers additional therapy to the patients who have a high-risk of recurrence suggested by adverse pathological features found in the surgical specimens [e.g. positive surgical margin (PSM), seminal vesicle invasion (SVI), EPE and higher Gleason scores (GS) 8 – 10], and/or other clinical features without evidence of disease recurrence (i.e. detectable PSA of 0.2 ng/ml). Salvage therapy may also be recommended to patients, which is the administration of treatment to patients with

biochemical recurrence (i.e. PSA level > 0.2 ng/ml with a second confirmation), persistent PSA, and/or local recurrence with/without evidence of metastatic disease. For patients without distant disease, the treatment is primarily via administrating radiation therapy with or without combining other treatments (e.g. hormone therapy) [12].

Selecting patients for adjuvant/salvage therapy after surgery is essential since adjuvant therapy may result in overtreatment for the patients. It may treat the patients who never would have developed recurrence. However, salvage therapy may result in the progression of the disease because its use is limited to the recurrence patients, particularly with high-risk disease [12].

Generally, adjuvant and salvage therapy have shown improved outcomes for RP patients. Patients with adverse features may benefit from adjuvant therapy and patients of certain risk groups [e.g. PSA doubling time (PSADT) of 6 months, PSM, pathological stage (pT) 3 cancer, GS 8 – 10] may benefit from salvage therapy. However, selecting patients for adjuvant or salvage therapy is still difficult since there is no certain conclusion to suggest that adjuvant or salvage therapy can improve the clinical outcomes for the specific patient group (e.g. GS 8/EPE alone). As a result, the literature [12] has noted the clinical need for evidence-based risk stratification with specific pathological features to advise adjuvant therapy. The benefit of salvage therapy may be specific to certain patient groups for a specific outcome. However, those findings were primarily based on observational studies and the benefit of salvage therapy was demonstrated by two randomized controlled trials [14, 15] that were based on internal subgroup analysis [12].

1.1.2.3 Pathology role for post-surgery patient care

Based on the treatment options and the benefits to the patient groups as discussed above, we can summarize that the adverse pathological features on RP specimens (i.e. PSM, EPE, SVI and GS 8 – 10) are important indicators for recurrence prediction and selection and guidance for adjuvant and salvage therapy.

In addition, pathological features derived from RP specimens (i.e. Gleason score, pathological stage, and tumour volume) have been demonstrated to have high prognostic value to predict disease progression [16-18], in which the Gleason score of the RP specimen is the most powerful predictor [19, 20]. Pound et al [13] found in their study that patients with surgical specimens of GS 8 – 10 usually were found having metastases within five years and those with GS 5 – 7 usually were found having metastases within ten years. A large retrospective study [21] has shown that GS 8 – 10, pre-external beam radiation therapy PSA level > 2 ng/ml, SVI, negative surgical margins, and PSADT ≤ 10 months were predictors of progression for 501 patients with biochemical recurrence who received salvage radiation therapy.

Pathological features may also have a role for potentially more specific patient stratification for guidance and selection for adjuvant and/or salvage therapy. Since the trials [14, 22, 23] studying the effectiveness of adjuvant and salvage therapy were not designed for subgroup analysis, the subgroup results have limitations because of 1) the inconsistency across trials in subgroup selection for analysis and findings across subgroups, and 2) insufficient statistical power since the initial design of those trials were not for subgroup analysis; for example, the stratifications were not randomized by subgroups. Although the results should be interpreted with caution due to those

limitations, the findings can be used as direction for future research. It is noted that pathological features, including positive/negative surgical margins, presence/absence of SVI, presence/absence of EPE, and Gleason scores, may be further analyzed for efficacy outcome and the clinical need for evidence based risk stratification to guide adjuvant therapy for patients with specific pathological findings [12].

1.1.3 Histopathology assessment on RP tissue sections

Histopathology for PCa refers to the assessment of the tissue under the microscope by pathologists for finding and studying the manifestation of disease. The tissue samples may come from biopsy or surgical specimens (RP specimens). The pathological assessment of prostate tissue for PCa is primarily on hematoxylin and eosin (H&E) stained tissue samples, which includes finding and grading the adenocarcinoma using the Gleason grading system, assessment of tumour size, tumour zonal origin, and the pathological stage. It also includes the identification of cancer penetrating through the prostate which includes EPE, PSM, and SVI. EPE means that the cancer has grown outside of the prostatic capsule into the peri-prostatic regions. PSM means that the cancer is found at the boundary where the surgeon's knife cut the tissue to remove the prostate. SVI means that the cancer has grown into the seminal vesicles. Some other features such as high-grade prostatic intraepithelial neoplasia (PIN), atrophy, and benign prostatic hyperplasia (BPH) may be noted in some clinical scenarios and used for study purposes. The subtypes of PCa are used for grading purposes, and are usually not reported independently in clinical practice. However, they may be reported for study purposes since some studies have shown the prognostic value of some subtypes.

This thesis primarily focuses on identifying, locating, and grading adenocarcinoma on RP specimens after surgery; thus, related background will be discussed in more detail (i.e., Gleason grading system, tumour volume, tumour zonal origin, and pathological stage on RP specimens). However, some histopathological biopsy related content will be discussed briefly for comparison purposes.

1.1.3.1 RP specimen processing protocol

RP specimens can be processed by either whole-mount or routine section [24]. After sectioning the apex and base of the removed specimen, the rest of the specimen is sectioned transversely at 3 to 5 mm intervals in a serial fashion. Routine sectioning refers the cutting of the transverse sections into quadrants to fit standard cassettes, while the whole-mount section refers to the intact transverse section. The tissue sections are then further processed by recutting and mounting on to the physical slides for staining (Figure 1.1) [25].

Although whole-mount sections are more expensive, harder to make, difficult to use for immunohistochemistry, and do not fit into standard slide holders for slide archives and collections, they yield a better overview for the pathologists and facilitate the identification of multifocal tumours. The primary advantages of using whole-mount sections are in displaying the prostate architecture, identifying and locating tumours more clearly [24]. These advantages support more straightforward tumour volume estimation [25], pathological staging, tumour origin zonal estimation and Gleason scoring [26]. In addition, technicians who are experienced in cutting whole-mount sections may find that it is less time consuming for them to cut whole-mount sections than cutting multiple blocks [24].

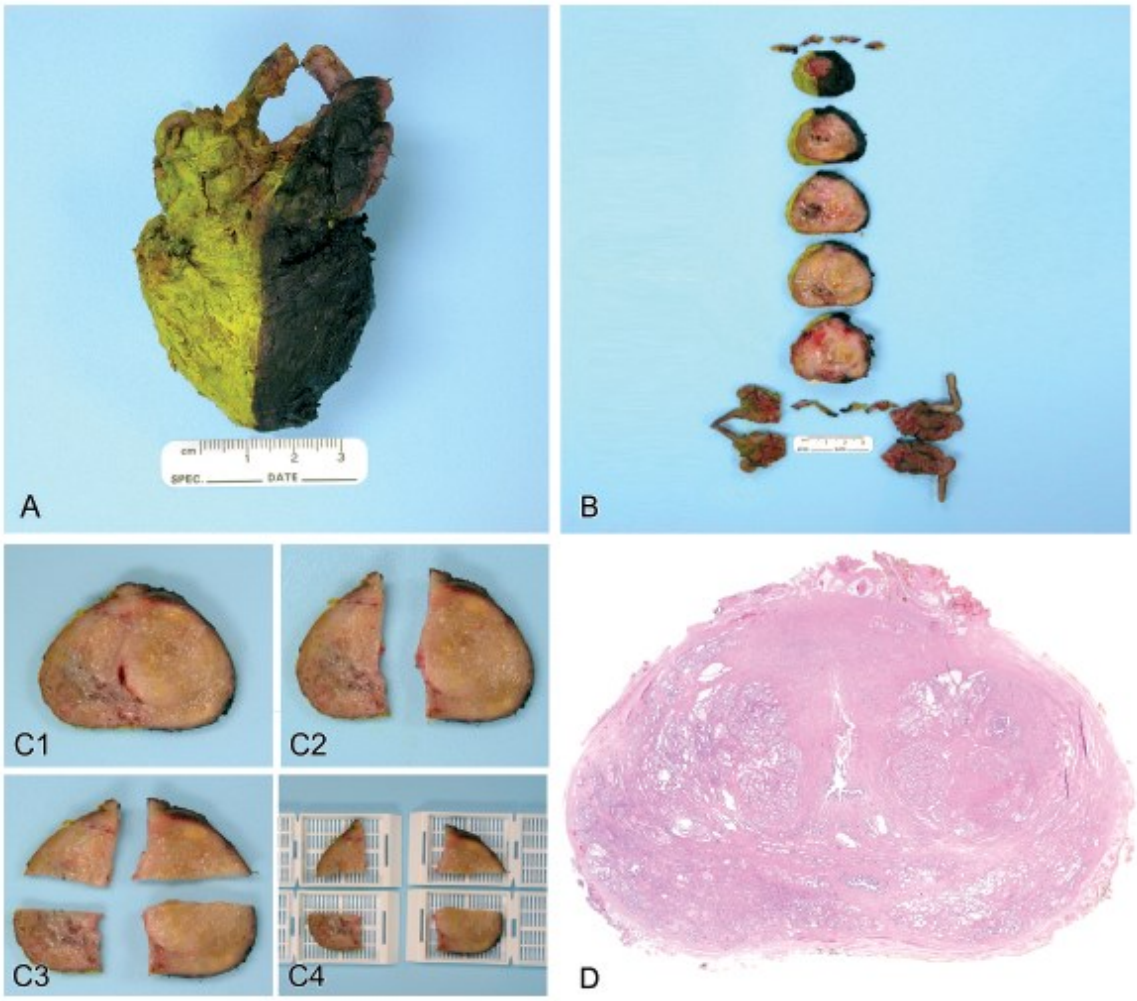


Figure 1.1: RP Specimen handled by whole-mount vs. routine section. Reproduced with permission from Sung et al. [25]. A: inked RP specimen. B: sections of RP. C1: whole-mount section. C2 – C4: sectioning whole-mount section into quadrant routine sections. D: H&E stained mid-gland whole-mount section.

Considering the advantages of using whole-mount sections, we used whole-mount sectioning in our study and the protocol is described in the following paragraph.

After surgery, we used 10% formalin buffered fixative to fix the specimen for 48 hours. The prostatic apex and base were removed and the mid-gland was sliced into 4.4 mm thick tissue sections which were then embedded using paraffin. The 4 μm thick tissue was taken by recutting the paraffin embedded tissue sections using the microtome and then mounted onto glass slides to be stained with H&E, which yields the H&E stained physical slide for pathological examination.

1.1.3.2 Gleason grading system for PCa

The Gleason grading system for PCa, which stratifies the aggressiveness of the tumours based on histological assessment, is a powerful tool for prognosis and aiding the treatment of PCa patients. It is the predominant prostate cancer grading method in research and daily practice worldwide. Its prognostic value was demonstrated by a study with long-term follow-up using survival as an endpoint in a large population [27].

The Gleason score, which sums the primary and secondary patterns [1] of the tumour on the RP specimen is the strongest indicator of tumour progression after RP [27].

1.1.3.2.1 Discrepancies in histology assessment between using biopsy and RP specimens

Biopsy is widely used in clinical practice for PCa diagnosis before and after treatment. However, pathology reporting from biopsy may differ from that of the RP specimens. The exact correlations in Gleason scores ± one Gleason score are in 43%, and 77% of cases respectively based on the data from 18 studies [17], where under-grading and over-grading in biopsy were respectively reported in 42% and 15% of reviewed cases [27]. In addition, using biopsy to determine histological boundaries of the tumour is

challenging unless the sampling is extremely dense, which makes it clinically impractical [28].

The source of the discrepancies is mainly from the heterogeneity of the PCa tissue and the small amount of the tissue being sampled by the biopsy cores (e.g. approximately 0.04% of the average gland (40 ml) was sampled by using 20 mm 18-gauge core biopsy samples), which results to sampling error, error in pathologic interpretation, and error from borderline cases. More specifically, the sampling error may result in under-grading when the sampling failed to sample the tissue with higher Gleason grade [29], and over-grading when sampling tissue with high-grade which only represents a very minor portion of the RP specimen. The error in pathologic interpretation of under-grading the biopsy core by the pathologist is common because the presence of confusing patterns changes the designation of the Gleason grade. For example, the limited foci of small glands of cancer may lead to the designation of G3 by definition. It is usually present in closely packed fashion instead of an infiltrating growth pattern. Challenges in recognizing the infiltrative growth pattern or presence of small areas of gland fusion (signs of presence of G4) from the biopsy cores may lead to under-grading [30]. In addition, for borderline cases (e.g. scoring a case of GS7 = G3+4 vs. GS7 = G4+3), intra- or inter- observer variability may be the source of discrepancies [27]. As a result, pathological features based on the assessment of RP specimens, when available, are considered essential for predicting the progression after RP.

1.1.3.2.2 Gleason grading system for RP specimens

The Gleason grading system was first proposed by Donald Gleason et al. [31] in 1966. It stratifies PCa into five grades reflecting its aggressiveness numerically based on

the microscopic morphological patterns of the tumour growth at lower power (i.e. 40 to 100X). The Gleason grade is increasing in numbers referring to the cancer growth: Gleason grade 1 (G1) to G5 range from well-differentiated to poorly-differentiated tumour tissue (Figure 1.2) [1]. Due to the heterogeneity of PCa, reporting the primary and secondary tumour (based on tumour size) grades is typically done, adding the grades to yield the Gleason score (GS). The system was first developed by evaluating biopsies, transurethral resections, and RPs from 270 PCa patients and tested on 1032 patients using cancer-related mortality as outcome. Since then, the system has been widely accepted and referred to as the Gleason grading system [17].

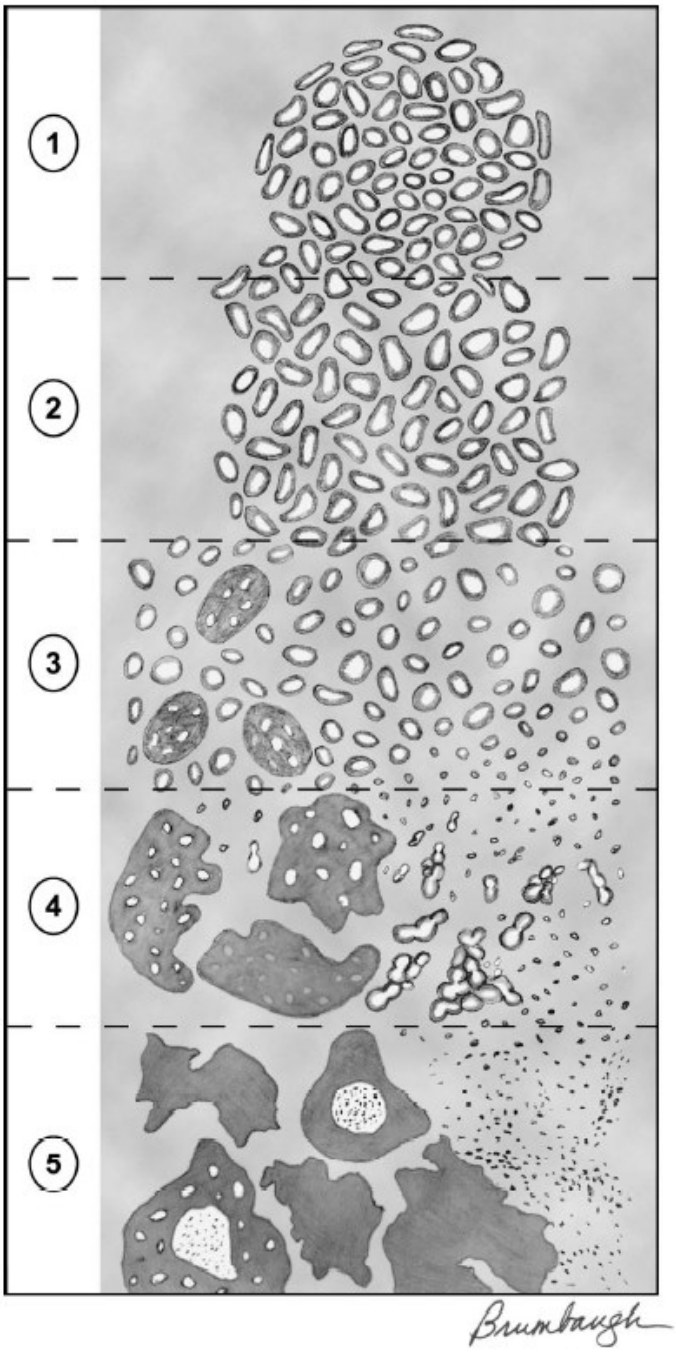


Figure 1.2: Illustrated Gleason patterns from the updated Gleason grading system. Reproduced with permission from Epstein et al. [32].

Since first proposed in 1966, the Gleason grading system has been influencing clinical practice. However, there have been many new discoveries and changes in clinical practice and pathology that justified the need for revision. The most recent revision was made by the International Society of Urological Pathology (ISUP) in the consensus meetings in 2005 and 2014. The resulting revised grading system is widely accepted and used [17].

The revision primarily made changes in the following aspects: 1) to reassess the G1 and G2 pattern, many of which are now considered as cancer mimickers instead of cancer; 2) to deal with the situation when high grade cancer is present as neither the primary nor the secondary tumour on RP, such that the presence of high-grade cancer may be noted as tertiary or replace the secondary grade based on the percentage of the high-grade cancer; 3) to address the issue of multifocal nodules on RP via recommending assigning Gleason score/grade group to each separate tumour on RP; 4) to incorporate newly discovered patterns [17].

In addition to those changes in the grading system, the newly developed grade group was first presented at the ISUP 2014 meeting and was widely accepted to replace the old 3-tier grade group system. The grade group was developed for patient risk-level stratification based on the GS to support treatment planning. The new system was initially developed in 2013 on a cohort of 7,869 RP patients at the John Hopkins Hospital and then validated on 20,845 patients from 5 academic centres [33]. The new grade group system stratifies patients into five groups reflecting risk-levels, based on the Gleason score. The five grade groups are: grade group 1 (GS6 = G3+3), grade group 2 (GS7 = G3+4), grade group 3 (GS7 = G4+3), grade group 4 (GS8 = G4+4), grade group 5 (GS9 –

10). In comparison, the old system frequently combined Gleason scores into a 3-tier groups of GS6, GS7, GS8 – 10 for prognosis and therapy guidance [17].

The major changes in the revision arise from the emphasis on the presence and the percentage of the high-grade cancer (i.e. G4 and G5) and dealing with multifocal specimens. Comparing to the old system, the new system separates the GS7 into two groups of G3+4 and G4+3, in which the G4 is secondary and primary grade respectively, due to the significant difference in prognosis found in the studies. Similarly, GS8 – 10 was separated into two groups of GS8 and GS9 – 10. The latter group indicates the presence of G5 and has much worse prognostic outcome [34]. In addition, for RP specimens, reporting the Gleason score based on the percentages of the presence of G4 and G5 is proposed with a threshold of less than 5% to be considered as tertiary. Otherwise, the GS is to be calculated by summing the primary Gleason grade and the highest Gleason grade. It is important to identify the percentage of G4 of very limited, close to 50%, or closer to 90%, which are used as borderlines for differentiating GS6 and GS7 = G3+4, GS7 = G3+4 and GS7 = G4+3, and GS7 = G4+3 and GS8, respectively [17]. In addition, a larger percentage of G4 is related to higher risk of biochemical failure after RP [35].

Although the revision noted the need for reporting the percentage of G4 and G5 on RP specimens to support better pathological reporting, it is challenging to report appropriately. Reporting on borderline cases of GS7 = G3+4 and GS7 = G4+3 based on the percentage of G4 is a chance of flipping coins [17], therefore can result in large inter- and intra-observer variability. Similarly, for reporting on borderline cases of GS6 versus GS7 = G3+4, the challenge remains, and it is recommended to include the percentage of

G4 for the purpose of confirmation. There is evidence that GS8 = G4+4 disease with the presence of G5 has different prognostic outcomes, comparing cases without G5 [33, 36]. However, it is still unclear whether GS8 = G3+5 and GS8 = G5+3 have different prognoses. Although one study [33] suggested that GS8 = G3+5 and GS8 = G4+4 have similar cancer-specific mortality rates, and GS8 = G5+3 and GS9 have similar cancer-specific mortality rates, it was noted that those results may be affected by potential error that may result from underscoring the separate nodules of GS10 and GS6 as an overall tumour of GS8 = G5+3 in an RP specimen [17].

1.1.3.3 Subtypes of PCa

The Gleason grading system classifies tissues based on the relationships of their morphological appearances to different prognostic outcomes. There is more than one underlying pattern in each of the Gleason grades. Following the most recent consensus from the ISUP meeting 2014 [34], the general rule for grading PCa is to grade well-formed glands as G3, cribriform, poorly formed, and fused glands as G4, and the absence of gland formation and necrosis as G5 [17]. More specifically, G3 includes the subtypes of sparse G3, intermediate G3, and packed G3, which depict well formed cancerous glands intervened with decreasing amounts of stroma tissue [37]. G4 includes the subtypes of "1) cribriform glands (including the glomerulid pattern), 2) poorly formed glands, and 3) fused glands." G5 includes the subtypes of "1) solid nests, 2) cords of cells, 3) individual cells, or 4) nests of cribriform glands with unequivocal necrosis [17]." Based on the consensus reached in 2014 ISUP meeting [34], mucinous adenocarcinoma should be graded based on the underlying pattern instead of designating it as independent grade of mucinous adenocarcinoma without the presence of extraglandular mucin, since

there is no necessary correlation between this type of cancer and an aggressive clinical behavior [17].

The subtypes are not limited to be used for grading the tissue using the Gleason grading system; literature has shown evidence that certain subtypes can be used as predictors for prognosis [38, 39]. For example, Dong et al. [39] found in their study that the presence of a cribriform pattern is predictive for biochemical recurrence and metastasis after RP. Also, Trudel et al. [38] found that the presence of any amount of large cribriform/intraductal carcinoma is a significant prognostic factor for estimating biochemical recurrence-free rate. However, these findings are limited since there are very few studies on this topic and larger scale studies using other clinical outcomes have been urged to integrate those findings into routine pathology practice [38]. In addition, the prognostic power of other subtypes were not discovered. It is challenging to conduct further study since detailed subtype annotation will be needed in large cohorts. This is not feasible within standard clinical pathology procedures, and requires significant extra effort from pathologists.

1.1.3.4 Tumour volume of RP for prognosis

Although most studies have failed to provide evidence that tumour volume and percentage of cancer involvement on RP have independent prognostic value, these parameters have been found to be associated with other pathological features (e.g. GS and EPE) [26]. Studies [32] have demonstrated that tumour volume is a predictive parameter for predicting the development of metastasis, SVI, and EPE. Some studies have shown that the percentage of cancer involvement has an even stronger correlation to pathological stage and tumour progression [40, 41]. Therefore, the prognostic

significance of providing quantitative information of tumour volume and proportion of tumour volume involvement is not disputed [26].

Reporting tumour volume in PCa is challenging since the assessment of tumour volume is technically difficult and is much more difficult than that for most other organs, and the methods for measuring PCa volume on RP specimens vary. The methods range from using computer assisted systems to naked eye examination without tumour annotations on the glass slides [26]. The methods using glass slides are more appropriate for routine clinical practice. The more sophisticated methods (e.g. counting numbers of involved blocks [42]) may involve much more effort from the pathologists, while the simpler methods (e.g. naked eye examination without tumour annotation on the slide [43]) is more subjective, which leads to intra- and inter- observer variability, especially considering the volume estimation requires virtual 3D reconstruction using 2D serial sections that are 3–5 mm apart and estimating tumour sizes on each 2D section. In addition, in the landmark studies [44-47], authors have noted that the difficulty of assessing the largest tumour on RP specimens. A method of measuring the maximum diameter of the largest tumour after annotating the tumour on the slide was proposed [48] as a surrogate for reporting.

In general, consensus was reached that it is required to report some quantitative measurement of tumour volume instead of using a qualitative description (e.g. reporting small vs. large tumour) with methods which are feasible in routine clinical practice based on each laboratory because of the prognostic value and the potentially superior importance for both clinical and pathological staging. However, the vote was nearly evenly split on reporting the maximum diameter as the clinical standard, which may be

due to the uncertainty of its independent prognostic value. Imaging techniques may support the need for reporting tumour size parameters and other parameters which reflect tumour volume [26].

1.1.3.5 Zonal origin of RP for prognosis

Most studies have shown that tumours originating from the peripheral zone are more aggressive than those from the transition zone, which usually have lower GS and pathological stage [26]. Transition zone tumours are more likely to have longer biochemical failure-free interval for recurrence patients after curative therapy, compared to peripheral zone tumours [49-52]. In addition, even for tumours from the transition zone having larger tumour volumes with significantly higher PSA levels than tumours from the peripheral zone, the biochemical cure rates are similar [53]. Although there are other studies [54, 55] showing contradictory findings, in a study analyzing zonal location of PCa as a possible indicator for progression-free survival post-surgery, Augustin et al [56] indicated that tumours from the transition zone, which is not an independent predictor in the multi-variate analysis, are correlated with better biochemical cure rate after RP [26].

Reporting the zonal origin in routine practice was not finalized and consensus was not reached, while consensuses were reached on reporting zonal location of the dominant/index tumour and dominant tumours found in the anterior of the gland. This might be due to the fact that the zonal origin is technically hard to determine and report on standard sections of RP specimens (quadrant sections from whole-mount). It is common to find that tumours cross both the transition and peripheral zones. In such

cases, the Gleason grade and the proportion of tumour in the transition zone may be used to deciding whether the tumour is from transition zone [26].

1.1.3.6 Pathological stage

Pathological stage, which stratifies PCa based on the findings reflecting tumour size and location on the sampled tissues, is considered as essential information for patient risk management. The final pathological stage from assessing the RP specimen is a more accurate predictor for cancer recurrence after surgery than other clinical preoperative parameters (e.g. PSA level, biopsy GS etc.). The staging criteria are summarized in Table 1.1 based on the most recent tumour-node-metastasis (TNM) system [57].

The debate regrading reporting the substages (i.e., pT2a, pT2b, and pT2c) of pT2 (organ confined) PCa was discussed in the 2009 ISUP consensus meeting due to the lack of evidence of prognostic value of the substages. Also, there are no uniform criteria and methods for staging pT2a and pT2b disease, which were defined as the unilateral cancer involvement less than and more than half of one lobe, respectively, which leads to technical challenges in reporting (e.g. large intra- and inter-observer variability). Consensuses were reached on discontinuing reporting the substages of pT2 using the 2002/2012 TNM guidelines. Recommending optional reporting of this parameter is under debate as most pathologists indicated a belief in its clinical and academic relevance in the pre-meeting survey [26].

Table 1.1: Pathological stage for prostate cancer. Reproduced with permission from AJCC **[57].**

Pathological stage (pT)*	
pT2	Organ confined
pT2a	Unilateral, one-half of one side or less
pT2b	Unilateral, involving more than one-half of side but not both sides
pT2c	Bilateral disease
pT3	Extraprostatic extension
pT3a	Extraprostatic extension or microscopic invasion of bladder neck**
pT3b	Seminal vesicle invasion
pT4	Invasion of rectum, levator muscles, and /or pelvic wall

* *Note*: There is no pathologic T1 classification.

** *Note*: Positive surgical margin should be indicated by an R1 descriptor (residual microscopic disease).

1.1.3.7 Role of computational tools for RP specimen pathological reporting

Although the importance of reporting tumour volume in pathological reports is widely accepted, the required substantial extra effort from the pathologists and the technical difficulties of assessing the tumour volume result hesitation to report tumour volume in clinical practice. The technical difficulties include inter- and intra- observer variabilities, and the lack of standard protocol of reporting. Digital pathology imaging techniques may enable automatic contouring of each individual tumour to have a graphical pathological report with associated tumour quantification. This could potentially yield more accurate tumour volumes, or estimation of the maximum diameter of the dominant tumours. Therefore, computational tools are expected to be essential to addressing questions around reporting of those parameters in routine clinical practice by supporting studies of the predictive power of those parameters for clinical outcomes [26]. Computational pathology systems may also be beneficial to more efficient pathological workflow for reporting zonal locations of dominant tumours and pT staging. In addition, the uncertainties related to reporting the pathological features (such as reporting maximum diameters and substage of pT2 PCa) may be resolved by studies using large cohorts with the graphical and quantitative information for each tumour on whole-mount RP sections provided by the computational pathology system.

1.1.4 Digital and computational pathology

The advent of digital pathology systems and increasing applications using machine learning in medical imaging analysis enable the use automated tools for analyzing pathology images to potentially assist clinical work and medical studies.

1.1.4.1 Digital pathology

Digital pathology is an image-based platform for pathological information acquisition, management, and interpretation from digitized slides. It is powered by virtual microscopy, which is a method for glass slide digitization and using a computer to review the digitized tissue of the glass slides [58]. Because of the progress in scanning technology, whole-slide-imaging (WSI) systems are available, which enable the scanner to scan the entire slide and save as a digital file for review. By using a modern WSI system, the pathologist can review a virtual slide in a way that is similar to reviewing a Google map [i.e. to navigate the virtual slide, and zoom in and out the regions of interest (ROIs) to review it at multiple resolutions] [59].

Compared to using the conventional physical slides, the use of WSI systems has advantages in terms of image viewing, remote consulting, portability, data archiving and retrieving, education, and integration of automated tools. The digitized files can be shared through the Internet to support distance teaching and remote pathology consultation [58]. Also, virtual slides are saved and managed electronically, they do not deteriorate or bleach over time and the accessibility of the file is higher. In addition, WSIs make the use of computational tools possible, for example using machine learning based algorithms to analyze the data (e.g. digital images, features extracted from the digital images etc.) from WSIs to assist with the diagnosis of cancer.

In order to use virtual slides for pathological examination, the slides have to be scanned at adequately high resolution and sufficient colour depth [59]. The resolution is usually represented using the unit of micrometer per pixel. For example, a virtual slide generated at 0.5 μm/pixel means the side length of each pixel in the image of the virtual

slide is equal to 0.5 μm in the physical slide. The colour depth is typically expressed using the number of bits per pixel. For example, a 24-bit colour depth in red-green-blue (RGB) colour image usually uses 8 bits for each of the R, G, B, such that it gives 16,777,216 colour variations. This 24-bit colour depth is also called “true colour”, which covers the range with more detail than human eye can perceive using a typical display [60].

Due to the high resolution with the given colour depth of the virtual slide, a single WSI can have a large file size, ranging from hundreds of megabytes for a single small region to multiple gigabytes for a single WSI of a tissue section. This raises challenges for viewing the scanned virtual slide, especially for reviewing through the Internet since a large amount of data needs to be rendered for the purposes of display and navigating. The solution to this challenge is the use of a pyramid structure to save the scanned virtual slide, which saves it at multiple resolutions (Figure 1.3). When a viewer needs to review a larger field of view, the computer can render the image by retrieving image data at much lower resolution. For example, using an Aperio Scanscope whole slide scanner to acquire a virtual slide at 20X, the same image is usually downsampled to 5X, 1.25X, and 0.625X respectively. This enables fast browsing and navigation of a large field of view of the image in a zoomed out view via rendering the image information through retrieving at low resolution (e.g. 1.25X) [59].

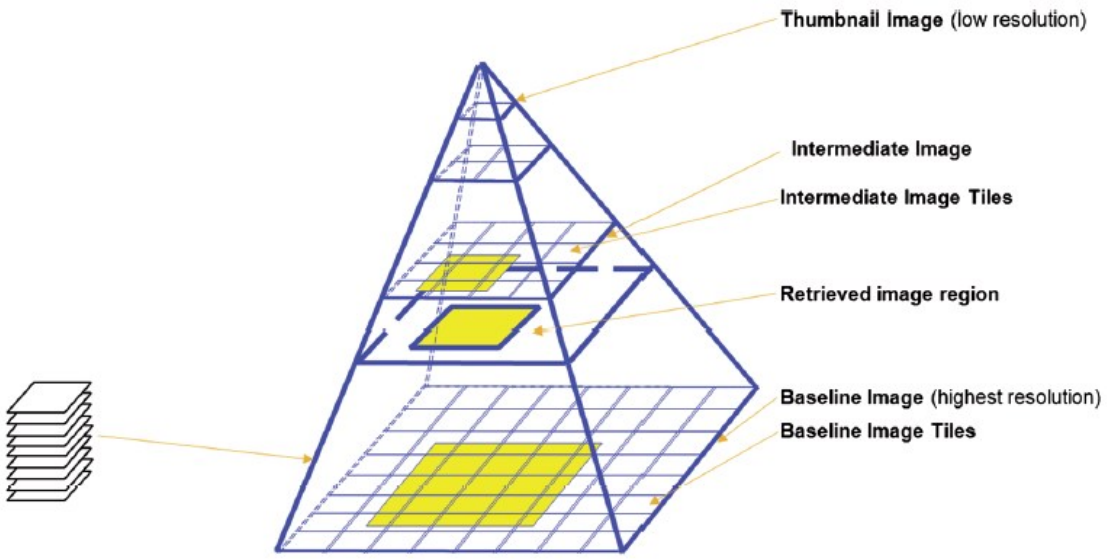


Figure 1.3: Pyramid structure of the digital pathology image file. Reproduced with permission from Higgins et al. **[58]**.

1.1.4.2 Computational pathology using machine learning

There is a more than two decade (since 1998) history of using computational methods to analyze pathology images to assist with disease diagnosis, prognosis, and quantitative characterization of biological features of the tissue. Machine learning plays important roles in pathology image analysis for, but not limited to, disease detection, classification, and object segmentation. Based on the machine learning methods used, it can be classified into two major categories: the traditional non-deep learning based methods and the deep learning based approaches. The non-deep learning based approaches include developing algorithms for object segmentation, detection, registration, and classification. For example, this approach can involve using feature engineering (e.g. extracting features from the images/sub-regions and converting to a

quantitative representation, selecting and analyzing features etc.) and/or machine learning methods (using data to build a model for prediction) to classify ROIs from the WSI as cancerous vs. non-cancerous [61]. The deep learning based approaches are one type of machine learning approach, which usually directly uses the "raw data" from the input (ROI of images/sub-images) without feature engineering and/or with minimal preprocessing.

Computational pathology has been impacted by deep learning. Deep learning based methods usually do not require engineered features. However, a large labeled training set is usually recommended to build such models/systems. Studies have demonstrated excellent performance using deep learning based methods for prostate cancer detection on digital pathology images [62, 63], outperforming traditional methods [61]. In addition, the deep learning based approaches are considered more flexible, which do not need intensive changes to the algorithms, while conventional machine learning methods usually require substantial efforts for feature engineering and data preprocessing. However, deep learning methods are known to require large amounts of annotated data [64, 65] and there is a noted difficulty in relating the results to the input for interpretation [66]. Therefore, estimating the robustness and generalization capability of such systems is challenging.

Given the evidence of the excellent performance of deep learning methods and the long history of traditional methods [61], it is essential to identify the application cases for each of the methods via direct comparisons using properly designed validation studies. This can help researchers and users to better understand the underlying principles of the deep learning based methods, therefore better estimating their robustness and

generalization capability, while on the other hand helping the algorithm design by integrating the findings (e.g., identified important features) from non-deep learning methods. Computational pathology is more and more widely accepted as key element for precision medicine, supporting quantitative and graphical pathology reporting and the development of computational pathology-based biomarkers for outcome prediction. Example use cases included as tumour volume estimation for PCa, and better treatment guidance (e.g., reporting the percentage of G3 and G4 cancer tissue for differentiating grade group 2 and 3 for more specific patient stratification for treatment recommendation).

1.2 Research challenges and related works

The work in this thesis includes computational pathology in the context of analyzing histological tissue of the prostate for automatic cancer detection and grading on whole-mount RP tissue sections. Many studies (see related studies in review [61]) have used computational methods to tackle problems of PCa detection and grading on H&E stained histopathology tissues in different aspects (e.g. using different methods/materials, aiming for different clinical purposes) [67, 68]. They demonstrated the feasibility and need for using computational pathology to detect and grade PCa on histology.

1.2.1 Computational pathology for PCa detection

Many studies [2, 62, 63, 67-84] have worked on classifying prostate tissue as cancer vs. non-cancer using computational methods. Based on the materials used in those studies, we can divide those studies into two broad categories of problems: 1) identification (i.e. given an ROI, indicate whether or not it is cancerous), and 2) detection

(i.e. given an image, indicate whether or not the image contains any cancer, and, optionally, indicate the locations of the cancerous regions). For the selected ROI based problem, studies usually worked with pre-selected ROIs with a smaller sample size [2, 63, 67-71, 73-76, 83, 84]. Those studies provided valuable insights in terms of potential methodologies, but the conclusions are limited by the small sample sizes used. Because prostate cancer tissue is heterogeneous, the generalization capability of the proposed systems in those studies is not clear. Also, whole-mount WSI files are usually large in size (e.g. 4~5 gigapixels, and each pixel may be 24-bits); thus, the efficiency of the system is important for clinical translation. Systems that were tested using pre-selected samples may not be scalable to process much larger tissue sizes within a practical time frame. In addition, there are large variations arising from tissue processing, therefore the robustness of the system is essential in application. Systems that were tested on samples from small patient cohorts may not be able to compensate for those variations.

Some studies (Table 1.2) developed and validated their methods on WSIs of prostate tissues. For the purpose of detection, studies usually worked with WSIs without sample selection. Most of the studies were conducted using biopsy tissues with the purpose of screening the negative samples to reduce the workload for pathologists. Since biopsy tissue is much smaller than RP tissue sections (e.g. sample tissues were approximately 0.04% of an average prostate gland, using a 20 mm long, 18 gauge biopsy needle core), the systems developed using biopsy tissues have similar limitations to those discussed above. In addition, biopsy tissues have sampling bias. Therefore, systems designed using those samples may not achieve the same performance on RP tissue sections. There are a few studies using WSIs of RP tissue sections [79-82]. In those

studies, only Monaco et al. [79] and Rashid et al [80] used whole-mount WSIs of RP tissue sections. In their studies, they noted limitations of their work in detecting high-grade cancer since the method is based on gland classification. Commonly, high-grade cancer may not have clear glands, as discussed in the background section, and is of high prognostic value clinically. In their studies, a very limited number of high-grade cancer samples was used. DiFranco et al. [81] used 15 WSIs from a 14-patient cohort without reporting system processing time. Because their methods extracted features at the pixel level at each of the colour channels, the processing time may be impractical for processing WSIs of RP sections. Nguyen et al. [82] reported a false positive rate of 6% and 78% sensitivity, testing on 11 WSIs. Because a very limited number of WSIs were used for testing in these two studies [81, 82], the generalization capability and practical applicability of the systems are not clear.

Table 1.2: A summary of previous work on PCa detection on WSIs. Acc.: accuracy. CV: cross-validation. LOO: leave-one-out. WM: whole-mount. AUC: area under the receiver-operating-characteristic (ROC) curve. The work described in this thesis is in boldface.

Year	1st Author	Results	Validation method	Data set	Total processed tissue (mm^2)	Tissue form	Reported speed (minutes)
2010	Monaco	0.87 (sensitivity), 0.9 (specificity)	3-fold CV	40 WSIs from 20 patients	14,000	RP	2.75/ WSI
2011	DiFranco	0.955 (AUC)	2-fold CV at WSI level	15 WSIs from 14 patients	10,134	RP	Not reported
2011	Nguyen	6% FPR, 78% TPR	2-way split	11 WSIs	316	RP	Not reported
2012	Doyle	0.84 (AUC)	3-fold CV at WSI level	100 WSIs from 58 patients	3,125	Biopsy	3/ROI
2016	Litjens	0.92 (AUC)	10-fold CV at patient level	204 WSIs from 163 patients	Up to ~4,800	Biopsy	4/WSI
2016	Litiens	0.99 (90th percentile) (AUC)	3-split at WSI level	225 WSIs from 50 patients	Up to ~5,600	Biopsy	Not reported
2019	Rashid	93% (Acc.)	2-split at WSI level	70 WSIs from 30 patients	Up to ~84,000	RP	Not reported
2019	**Han**	**0.98 (AUC) 94.4% (Acc.)**	**LOO, 5-fold, 2-fold CV at patient level**	**299 WSIs from 71 patients**	**358,800**	**WM RP**	**2 /whole-mount WSI**

In short, it is important to develop and validate systems using WSIs of whole-mount RP tissue sections from a large patient cohort with strict validation methods (e.g. stratifying training and test cases at the per-patient level) with practical processing time for translational purposes. Our work aimed to fill the gap between the previous work and this purpose.

1.2.2 Feature extraction

Features are typically calculated to characterize image samples for classification. At a high-level, the methods for calculating those features may be divided into two major streams: 1) extracting features directly from digital images at the pixel level, and 2) extracting features at the object level (e.g. tissue components) after object segmentation or identification.

Some studies used features that were calculated at the pixel level (see review in [61]). The primary advantage of those methods is the use of most of the information within the sample images. However, challenges include the resulting variations in calculated features that come from the image variability, which may result from various artifacts such as staining variability [85]. Also, the computational cost is usually high based on the reported processing times in previous studies. To resolve the issue of staining variability, normalization has typically been used [86-88]. Normalization methods usually used the selected images as standard/target images for calibration, or normalized pixel colour/intensity across selected image samples.

Extracting features at the object level usually requires accurate object segmentation and identification. For object segmentation, methods include using machine

learning to classify each object, or segmentation algorithms to segment objects directly [61]. However, staining variability is still a problem that negatively affects the segmentation results [61, 85]. This is because, similar to normalization as discussed above, classification used other image samples for the purpose of system training, and segmentation algorithms were usually performed after normalization. Studies were usually single centre studies with fewer than hundreds of manually annotated samples for training and validation.

In short, for sample image classification (e.g. cancer vs non-cancer or grading), extracting features that are robust to staining variability and fast to compute is still a challenging problem domain. Our work developed a tissue component segmentation method that is independent of other sample images, and is fast in computation to support consistent feature extraction.

1.2.3 Computational pathology for PCa grading based on the Gleason grading system

For grading PCa as high- vs. low-grade, studies (Table 1.3) used various methods, but those methods were similar to the methods used for classifying cancer vs. non-cancer at a high-level. They also typically used machine learning based approaches to classify positive and negative image samples after feature extraction. The challenges in PCa detection, as discussed above, are also applicable to this problem. In addition, grading PCa on histology is a more challenging problem because morphological patterns of cancerous tissues between different Gleason grades are more similar, compared to those between cancer and non-cancer tissues. For example, packed G3 and small fused G4 are more similar in morphology, and both show packed cancer glands with different amounts

of fused glands. Therefore, those tissues are harder to differentiate. Also, for samples having the same Gleason grade, there are various patterns. For example, G4 can be divided into cribriform G4, poorly formed G4, and small fused G4. The system needs to be able to identify all those patterns and correctly classify them as G4. In addition, PCa is highly heterogeneous, thus the cancerous tissues show large variations in appearance across different patients. Validation using samples from the same patient may positively bias the system performance. In short, it is important to design a grading system with high accuracy that was tested on all cancerous regions of WSIs of RP sections, which covers enough variability of cancerous samples for each Gleason grade. It is also important to validate the system using samples from different patients that were not used for training.

Table 1.3: A summary of selected previous work on PCa grading using computational methods. The listed works were selected based on the relevance of problem and methodology to this thesis. The work described in this thesis is in boldface.

Year	1st Author	Results	Validation method	Data set	Total processed tissue (mm^2)	Tissue form	Reported speed
2007	Naik	95.2% (Acc.)	2-fold	20 ROIs	Not reported	Not reported	Not reported
2012	Doyle	G3 0.77, G4 0.76, G5 0.95 (Acc.)	3-fold CV	2000 ROIs from 214 patients	0.13~1.3	Biopsy	Not reported
2013	Gorelick	85% (Acc.)	LOO CV	120 ROIs from 15 patients	2.7	RP	2 minutes/ROI
2013	Sparks	0.78 (AUC)	2-fold at patient level	120 ROIs from 58 patients	Not reported	Biopsy	Not reported
2014	Nguyen	0.82 (AUC)	LOO CV	221 cores of 368 patients	108	TMA	Not reported
2016	Niazi	90.9% (Acc.)	2-split	88 ROIs from 58 WSIs	178	Not reported	Not reported
2018	Nir	79.2% (Acc.)	LOO CV at patient level	333 cores from 231 patients	333	TMA	14 hours total
2019	**Han**	**0.92 (AUC)**	**LOO CV at patient level**	**286 RPs from 68 patients**	**17124**	**RP**	**<2 minutes/WSI**

In previous work [2, 83, 89-93], all studies used pre-selected ROIs and total processed tissue sizes ranged from 0.13 mm^2 to 333 mm^2. We observe from Table 1.3 that from the earlier work [2, 89, 92, 93] to more recent work [83, 90, 91], the reported system performances have not necessarily improved with time, and in fact some more recent studies [83, 90] reported worse performance than earlier studies [2, 93]. However, larger tissue amounts were used in the more recent studies. This suggests that grading was becoming more challenging when processing a larger amount of tissue, which brings greater variability of tissue patterns. The most recent work demonstrated the need for processing a large amount of tissue that covers large variability of tissue samples with more strict validation methods. Nir et al. [83] processed a total amount of approximately 333 mm^2 of tissue that consisted of 333 tissue microarrays (TMAs), which were sampled from 231 RP tissue sections. They tested their system used leave-one-out CV with data grouped on a per-patient basis and reported an accuracy of 79.2% for high- (G4 and G5) vs. low-(G3) grade classification with total processing time of 14 hours. Our work reported AUCs of 0.92 for classifying high-(G4 and G5) vs. low-(G3) grade and used all cancerous tissues from 286 RP sections (i.e. total amount of 17124 mm^2) without sample preselection, covering all clinically relevant grade groups, with a processing time of 2 minutes per mid-gland whole-mount WSI.

1.2.4 Technical methodology

As discussed above, a large number of publications demonstrated the feasibility of using computational methods to detect and grade PCa on digital histopathology images. Most studies used machine learning based approaches for classification after extracting features. Recently, deep learning based approaches have demonstrated their efficacy for

image classification problems [94]. Studies [62, 63, 83, 84] have also demonstrated excellent performance of using this method to analyze the digital histopathology images for PCa detection. Litjens et al. [62] reported an AUC of 0.99 (90^{th} percentile) using 238 biopsy tissues. Chen et al. [84] reported an AUC of 0.99 (95% confidence interval, 0.97-0.99) using 1360 ROIs from 34 WSIs. Kwak et al. [63] reported an AUC of 0.974 (95% CI:0.961–0.985) using 655 TMAs. Nir et al. [83] reported an overall accuracy of 91.9% using 333 TMAs from 231 patients. In a study [83], Nir et al. also used a deep learning based approach to grade cancerous TMAs as high- (G4 and G5) vs. low- (G3), and reported an overall accuracy of 77.8%.

Most studies using deep learning based approaches used raw image samples as input without pre-processing [62, 83, 84], while Kwak et al. [63] used both raw images and segmented nuclei maps as inputs for comparison. In addition, they also compared conventional machine learning based approaches to deep learning approaches. In their study, they found that the deep learning based approach using their nuclei seed maps as input yielded the best performance. Since deep learning is considered as a “black box”, lacking comprehensive understanding [95], it is more difficult to interpret the results and estimate the generalization capability and robustness of the system for practical application. Kwak et al. [63] demonstrated the need for and importance of comparing deep learning and conventional machine learning based approaches, and that of incorporating preprocessing into the deep learning based methods.

We summarize that existing studies using deep learning based methods used WSIs of biopsy specimens and TMAs. One of those studies compared the methodologies between deep learning and conventional machine learning, and incorporated image

preprocessing into the deep learning method. Our work used whole-mount WSIs of RP sections without sample selection for both PCa detection and grading. We provided comparisons between using conventional machine learning based approaches and using deep learning based approaches, and comparisons using 4 different tissue component maps and using raw image input.

1.2.5 Computational pathology for PCa sub-type grading

Although there are many studies in PCa detection and grading on histopathology images as discussed above, to the best of our knowledge, there is no study that used computational methods to grade PCa sub-types beyond Gleason grade. The need for this study was discussed in section 1.1.3.3. Our work used a deep learning based approach to detect each of eight sub-types on the WSIs of whole-mount RP tissue sections.

1.3 Thesis outline

To resolve the technology gap in applying machine learning methods for PCa detection and grading on whole-mount RP sections for clinical translation, the primary objective of this thesis work is to develop and validate an automatic system. The system needs to be accurate and fast. The validation needs to be done using all tissues from whole-mount RP sections, grouped on a per patient basis for training and testing.The major research questions are:

1. Can features extracted from 3-class (i.e. nuclei, lumen, and stroma/other) tissue component maps (TCMs) provide the major information for PCa detection and grading on whole-mount RP sections?

2. Can 3-class TCMs compensate for staining variability for robust PCa detection and grading on whole-mount RP specimens?
3. What is the most important information on the histology tissue that can be used for PCa detection and grading?
4. How do deep learning based approaches perform for PCa detection and grading?
5. What is the feasibility of detecting subtypes of G3 and G4 PCa using a deep learning approach?

To answer these research questions, this thesis breaks down the objective into the following specific aims using whole-mount RP tissue sections:

1. To develop an algorithm to segment tissue components from histology images and validate its performance for PCa detection and grading.

2. To develop and validate a machine learning based system using segmented tissue component maps for PCa detection and grading.

3. To validate and compare the performance using different tissue component maps (i.e. nuclei maps, lumen maps, and 3-class TCMs) for PCa detection and grading.

4. To develop and validate a deep learning based pipeline for PCa detection and grading and compare the performance to that using conventional machine learning based approaches.

5. To develop and validate a machine learning based system for PCa sub-type grading .

Chapter 2: Automatic cancer detection on digital histopathology images of mid-gland radical prostatectomy specimens

The objective of this work was to develop and validate an automatic system for PCa detection on whole-mount WSIs of H&E-stained RP tissue sections. We aimed for the developed system to be robust to staining variability across WSIs, efficient in computation, and accurate in performance with CV grouping data on a per patient basis using a large data set. This would support clinical translation upon further multi-centre validation and user-study. We generated 3-class TCMs using our proposed segmentation method, allowing fast and accurate tissue component segmentation. First and second-order statistical features were extracted at the object-level (i.e. from 3-class TCMs) to train three different classifiers to improve the generalization capability and robustness of the system. We used two deep learning based methods of transfer learning by fine-tuning pre-trained AlexNet (pre-trained by ImageNet data [64]) [94] with minimal preprocessing (i.e. down-sampling using binary interpolation to conform the input image size) of raw image samples and 3-class TCMs, respectively, for comparison. We demonstrated the state-of-art performance of the presented systems, and the fastest processing time comparing to that in literature, with validation conducted using the largest expert-annotated data set thus far. We found that 3-class TCMs encoded most the information in the form of tissue component patterns for PCa detection and grading, can compensate for staining variability, and reduced the sensitivity to sample size for deep learning based systems for accurate and consistent performance in PCa detection.

Chapter 3: Histological tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens

The purpose of this work was to grade cancerous regions on whole-mount RP tissue sections using an automated pipeline. We aimed to re-tune the pipeline that was developed for PCa detection as described in chapter 2, for PCa grading. We expanded previous pipeline by using nuclei maps and lumen maps respectively as system inputs for comparison. We aimed to discovery the importance of each of the 3-class tissue components for both PCa detection and grading. In addition, we comprehensively compared the conventional machine learning based approach and the deep learning based approaches for both PCa detection and grading, which includes direct comparisons in terms of overall performance and performance for each tissue type (e.g. G4 cancer, G5 cancer etc.). This will help us to better understand the underlying principles behind using a machine learning based approach for PCa detection and grading, and to estimate the utility of those methods. We found that the 3-class TCM included most of the information for both PCa detection and grading, and nuclei maps provided the key information for identifying high-grade (i.e. GS $\geqslant$ 9) cancer.

Chapter 4: Automatic prostate cancer sub-grading on digital histopathology images of radical prostatectomy specimens

The purpose of this work is to investigate the feasibility of using a machine learning based approach to automatically detect each sub-type of cancerous tissue on WSIs of whole-mount RP sections. We aimed to use transfer learning to classify all the cancerous samples as each sub-type vs. other. In addition, because deep learning demonstrated overall better performance than conventional machine learning based approaches when a large sample size is given, we used a deep learning based method for

this novel and challenging problem. Since a much smaller sample size is available for each of the sub-type, we used transfer learning to reduce the usual requirement of a large sample size for system training. We found that transfer learning that fine-tunes pre-trained AlexNet with raw image samples (i.e. image samples extracted from the WSI with minimal processing consisting only of downsampling), can be used for automatically detecting each of eight different sub-types of cancerous tissue beyond Gleason grade, and system performance is subject to sub-type.

Chapter 5: contributions and impact of the thesis, and future directions

This chapter summarizes the achievements, impacts, advances in knowledge relating to each of the research questions, potential applications, and future directions relating to the work of this thesis.

1.4 References

1. Gleason, D.F., G.T. Mellinger, and G. Veterans Administration Cooperative Urological Research, *Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. 1974.* J Urol, 2002. **167**(2 Pt 2): p. 953-8; discussion 959.

2. Gorelick, L., et al., *Prostate histopathology: learning tissue component histograms for cancer detection and classification.* IEEE Trans Med Imaging, 2013. **32**(10): p. 1804-18.

3. Levy, I.G., N.A. Iscoe, and L.H. Klotz, *Prostate cancer: 1. The descriptive epidemiology in Canada.* Cmaj, 1998. **159**(5): p. 509-513.

4. Canadian Cancer Society's Advisory Committee on Cancer Statistics*, Canadian Cancer Statistics* 2017, Toronto, ON: Canadian Cancer Society,2017.

5. Wong, Y.-N., et al., *Survival associated with treatment vs observation of localized prostate cancer in elderly men.* Jama, 2006. **296**(22): p. 2683-2693.

6. Bill-Axelson, A., et al., *Radical prostatectomy versus watchful waiting in early prostate cancer.* New England journal of medicine, 2005. **352**(19): p. 1977-1984.

7. D, B. and W. T, *Adjuvant radiotherapy after radical prostatectomy: indications, results and side effects.* Urologia Internationalis, 2007.

8. Stephenson, A.J., et al., *Predicting the outcome of salvage radiation therapy for recurrent prostate cancer after radical prostatectomy.* J Clin Oncol, 2007. **25**(15): p. 2035-41.

9. Gillitzer, R. and J. Thüroff, *Relative advantages and disadvantages of radical perineal prostatectomy versus radical retropubic prostatectomy.* Critical reviews in oncology/hematology, 2002. **43**(2): p. 167-190.

10. Walsh, P.C. and H. Lepor, *The role of radical prostatectomy in the management of prostatic cancer.* Cancer, 1987. **60**(S3): p. 526-537.

11. NCCN, *NCCN Clinical Practice Guidelines in Oncology Prostate Cancer 2019.* NCCN Guidelines, 2019.

12. Thompson, I.M., et al., *Adjuvant and Salvage Radiotherapy after Prostatectomy: ASTRO/AUA Guideline.* ASTRO/AUA Guideline, 2019.

13. Pound, C.R., et al., *Natural history of progression after PSA elevation following radical prostatectomy.* Jama, 1999. **281**(17): p. 1591-1597.

14. Bolla, M., et al., *Postoperative radiotherapy after radical prostatectomy for high-risk prostate cancer: long-term results of a randomised controlled trial (EORTC trial 22911).* The Lancet, 2012. **380**(9858).

15. Thompson, I.M., et al., *Adjuvant Radiotherapy for Pathological T3N0M0 Prostate Cancer Significantly Reduces Risk of Metastases and Improves Survival: Long-Term Followup of a Randomized Clinical Trial.* Journal of Urology, 2009. **181**(3): p. 956-962.

16. Van Der Kwast, T.H., et al., *International Society of Urological Pathology (ISUP) consensus conference on handling and staging of radical prostatectomy specimens. Working group 2: T2 substaging and prostate cancer volume.* Modern pathology, 2011. **24**(1): p. 16.

17. Kryvenko, O.N. and J.I. Epstein, *Prostate cancer grading: a decade after the 2005 modified Gleason grading system.* Arch Pathol Lab Med, 2016. **140**(10): p. 1140-1152.

18. van Oort, I.M., C.A. Hulsbergen-vandeKaa, and J.A. Witjes, *Prognostic Factors in Radical Prostatectomy Specimens: What Do We Need to Know from Pathologists?* european urology supplements, 2008. **7**(12): p. 715-722.

19. Roberts, W.W., et al., *Contemporary identification of patients at high risk of early prostate cancer recurrence after radical retropubic prostatectomy.* Urology, 2001. **57**(6): p. 1033-1037.

20. Epstein, J.I., et al., *Prognostic factors and reporting of prostate carcinoma in radical prostatectomy and pelvic lymphadenectomy specimens.* Scandinavian Journal of Urology and Nephrology, 2005. **39**(sup216): p. 34-63.

21. Stephenson, A.J., et al., *Salvage Radiotherapy for Recurrent Prostate Cancer After Radical Prostatectomy.* JAMA. **291**(11).

22. T, W., B. D, and B. D, *Adjuvant radiotherapy versus wait-and-see after radical prostatectomy: 10-year follow-up of the ARO 96-02/AUO AP 09/95 trial.* European Urology, 2014. **66**(2).

23. Swanson, G.P., et al., *The Prognostic Impact of Seminal Vesicle Involvement Found at Prostatectomy and the Effects of Adjuvant Radiation: Data From Southwest Oncology Group 8794.* Journal of Urology, 2008. **180**(6): p. 2453-2458.

24. Montironi, R., et al., *Handling of radical prostatectomy specimens: total embedding with large-format histology.* International journal of breast cancer, 2012. **2012**.

25. Sung, M.-T. and L. Cheng, *Contemporary approaches for processing and handling of radical prostactomy specimens.* Histology and histopathology, 2010.

26. van der Kwast, T.H., et al., *International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 2: T2 substaging and prostate cancer volume.* Mod Pathol, 2011. **24**(1): p. 16-25.

27. Montironi, R., et al., *Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies.* BJU Int, 2005. **95**(8): p. 1146-1152.

28. Gibson, E., *3D fusion of histology to multi-parametric MRI for prostate cancer imaging evaluation and lesiontargeted treatment planning.* 2014.

29. Algaba, F.A., et al., *Evidence of the radical prostatectomy Gleason score in the biopsy Gleason score.* Actas urologicas espanolas, 2004. **28**(1): p. 21-26.

30. Steinberg, D.M., et al., *Correlation of prostate needle biopsy and radical prostatectomy Gleason grade in academic and community settings.* Am J Surg Pathol, 1997. **21**(5): p. 566-576.

31. Gleason, D., *Histological grading and clinical staging of prostatic carcinoma.* Urologic pathology. The prostate, 1977. **171**.

32. Epstein, J.I., et al., *The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma.* Am J Surg Pathol, 2005. **29**(9): p. 1228-42.

33. Epstein, J.I., et al., *A contemporary prostate cancer grading system: a validated alternative to the Gleason score.* Eur Urol, 2016. **69**(3): p. 428-435.

34. Epstein, J.I., et al., *The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System.* Am J Surg Pathol, 2016. **40**(2): p. 244-52.

35. Sauter, G., et al., *Clinical utility of quantitative Gleason grading in prostate biopsies and prostatectomy specimens.* Eur Urol, 2016. **69**(4): p. 592-598.

36. Mahal, B.A., et al., *Gleason score 5+ 3= 8 prostate cancer: much more like Gleason score 9?* BJU Int, 2016. **118**(1): p. 95-101.

37. Downes, M.R., et al., *Determination of the association between T2-weighted MRI and Gleason sub-pattern: a proof of principle study.* Academic radiology, 2016. **23**(11): p. 1412-1421.

38. Trudel, D., et al., *Prognostic impact of intraductal carcinoma and large cribriform carcinoma architecture after prostatectomy in a contemporary cohort.* Eur J Cancer, 2014. **50**(9): p. 1610-1616.

39. Dong, F., et al., *Architectural heterogeneity and cribriform pattern predict adverse clinical outcome for Gleason grade 4 prostatic adenocarcinoma.* Am J Surg Pathol, 2013. **37**(12): p. 1855-1861.

40. Epstein, J.I., J.E. Oesterling, and P.C. Walsh, *Tumor volume versus percentage of specimen involved by tumor correlated with progression in stage A prostatic cancer.* J Urol, 1988. **139**(5): p. 980-983.

41. Partin, A.W., et al., *Morphometric measurement of tumor volume and per cent of gland involvement as predictors of pathological stage in clinical stage B prostate cancer.* J Urol, 1989. **141**(2): p. 341-345.

42. Jones, E.C., *Resection margin status in radical retropubic prostatectomy specimens: relationship to type of operation, tumor size, tumor grade and local tumor extension.* J Urol, 1990. **144**(1): p. 89-93.

43. Carvalhal, G.F., et al., *Visual estimate of the percentage of carcinoma is an independent predictor of prostate carcinoma recurrence after radical prostatectomy.* Cancer: Interdisciplinary International Journal of the American Cancer Society, 2000. **89**(6): p. 1308-1314.

44. Renshaw, A.A., H. Chang, and A.V. D'Amico, *Estimation of tumor volume in radical prostatectomy specimens in routine clinical practice.* American journal of clinical pathology, 1997. **107**(6): p. 704-708.

45. Eichelberger, L.E., et al., *Maximum tumor diameter is an independent predictor of prostate-specific antigen recurrence in prostate cancer.* Modern pathology, 2005. **18**(7): p. 886.

46. Renshaw, A.A., et al., *Maximum diameter of prostatic carcinoma is a simple, inexpensive, and independent predictor of prostate-specific antigen failure in radical prostatectomy specimens: validation in a cohort of 434 patients.* American journal of clinical pathology, 1999. **111**(5): p. 641-644.

47. Stamey, T.A., et al., *Biological determinants of cancer progression in men with prostate cancer.* Jama, 1999. **281**(15): p. 1395-1400.

48. Mai, K.T., et al., *A simple technique for calculation of the volume of prostatic adenocarcinomas in radical prostatectomy specimens.* Pathology-Research and Practice, 2003. **199**(9): p. 599-604.

49. Greene, D.R., J.M. Fitzpatrick, and P.T. Scardino. *Anatomy of the prostate and distribution of early prostate cancer*. in *Seminars in surgical oncology*. 1995. Wiley Online Library.

50. Shannon, B.A., J.E. Mc Neal, and R.J. Cohen, *Transition zone carcinoma of the prostate gland: a common indolent tumour type that occasionally manifests aggressive behaviour.* Pathology, 2003. **35**(6): p. 467-471.

51. Erbersdobler, A., et al., *Pathological and clinical characteristics of large prostate cancers predominantly located in the transition zone.* Prostate cancer and prostatic diseases, 2002. **5**(4): p. 279.

52. Noguchi, M., et al., *An analysis of 148 consecutive transition zone cancers: clinical and histological characteristics.* J Urol, 2000. **163**(6): p. 1751-1755.

53. Sakai, I., et al., *Analysis of differences in clinicopathological features between prostate cancers located in the transition and peripheral zones.* International journal of urology, 2006. **13**(4): p. 368-372.

54. Sakai, I., et al., *A comparison of the biological features between prostate cancers arising in the transition and peripheral zones.* BJU Int, 2005. **96**(4): p. 528-532.

55. Al-Ahmadie, H.A., et al., *Anterior-predominant prostatic tumors: zone of origin and pathologic outcomes at radical prostatectomy.* Am J Surg Pathol, 2008. **32**(2): p. 229-235.

56. Augustin, H., et al., *Zonal location of prostate cancer: significance for disease-free survival after radical prostatectomy?* Adult Urology, 2003. **62**(1).

57. AJJC, *AJCC cancer staging manual seventh edition.*

58. Higgins, C., *Applications and challenges of digital pathology and whole slide imaging.* Biotechnic & Histochemistry, 2015. **90**(5): p. 341-347.

59. Zarella, M.D., et al., *A practical guide to whole slide imaging: a white paper from the digital pathology association.* Arch Pathol Lab Med, 2018. **143**(2): p. 222-234.

60. Judd, D.B., *Color in business science and industry.* Applied Spectroscopy, 1953. **7**(2): p. 90-91.

61. Mosquera-Lopez, C., et al., *Computer-Aided Prostate Cancer Diagnosis From Digitized Histopathology: A Review on Texture-Based Systems.* IEEE Rev Biomed Eng, 2015. **8**: p. 98-113.

62. Litjens, G., et al., *Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis.* Sci Rep, 2016. **6**: p. 26286.

63. Kwak, J.T. and S.M. Hewitt, *Nuclear architecture analysis of prostate cancer via convolutional neural networks.* IEEE Access, 2017. **5**: p. 18526-18533.

64. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.

65. Shin, H.C., et al., *Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning.* IEEE Trans Med Imaging, 2016. **35**(5): p. 1285-98.

66. Madabhushi, A. and G. Lee, *Image analysis and machine learning in digital pathology: Challenges and opportunities*. 2016, Elsevier.

67. Hamilton, P.W., et al., *Automated histometry in quantitative prostate pathology.* Anal Quant Cytol Histol, 1998. **20**(5): p. 443-460.

68. Bartels, P.H., et al., *Machine vision in the detection of prostate lesions in histologic sections.* Anal Quant Cytol Histol, 1998. **20**(5): p. 358-364.

69. Diamond, J., et al., *The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia.* Human pathology, 2004. **35**(9): p. 1121-1131.

70. Tabesh, A., et al., *Multifeature prostate cancer diagnosis and Gleason grading of histological images.* IEEE Trans Med Imaging, 2007. **26**(10): p. 1366-78.

71. Farjam, R., et al., *An image analysis approach for automatic malignancy determination of prostate pathological images.* Cytometry B Clin Cytom, 2007. **72**(4): p. 227-40.

72. Tahir, M.A. and A. Bouridane, *Novel round-robin tabu search algorithm for prostate cancer classification and diagnosis using multispectral imagery.* IEEE Trans Inf Technol Biomed, 2006. **10**(4): p. 782-93.

73. Bouatmane, S., et al., *Round-Robin sequential forward selection algorithm for prostate cancer classification and diagnosis using multispectral imagery.* Machine Vision and Applications, 2011. **22**(5): p. 865-878.

74. Sun, X., et al., *Automatic diagnosis for prostate cancer using run-length matrix method*. SPIE Medical Imaging. Vol. 7260. 2009: SPIE.

75. Yu, E., et al. *Detection of prostate cancer on histopathology using color fractals and Probabilistic Pairwise Markov models*. in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2011.

76. Peyret, R., et al., *Automatic classification of colorectal and prostatic histologic tumor images using multiscale multispectral local binary pattern texture features and stacked generalization.* Neurocomputing, 2018. **275**: p. 83-93.

77. Doyle, S., et al., *A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies.* IEEE Trans Biomed Eng, 2012. **59**(5): p. 1205-18.

78. Litjens, G., et al., *Automated detection of prostate cancer in digitized whole-slide images of H and E-stained biopsy specimens*. SPIE Medical Imaging. Vol. 9420. 2015: SPIE.

79. Monaco, J.P., et al., *High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models.* Med Image Anal, 2010. **14**(4): p. 617-29.

80. Rashid, S., et al., *Automatic pathology of prostate cancer in whole mount slides incorporating individual gland classification.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2019. **7**(3): p. 336-347.

81. DiFranco, M.D., et al., *Ensemble based system for whole-slide prostate cancer probability mapping using color texture features.* Comput Med Imaging Graph, 2011. **35**(7-8): p. 629-45.

82. Nguyen, K., A.K. Jain, and B. Sabata, *Prostate cancer detection: Fusion of cytological and textural features.* J Pathol Inform, 2011. **2**: p. S3.

83. Nir, G., et al., *Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts.* Med Image Anal, 2018. **50**: p. 167-180.

84. Chen, P., et al., *An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis.* Nature medicine, 2019. **25**(9): p. 1453-1457.

85. Leo, P., et al., *Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images.* J Med Imaging (Bellingham), 2016. **3**(4): p. 047502.

86. Magee, D., et al. *Colour normalisation in digital histopathology images*. in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*. 2009. Daniel Elson.

87. Basavanhally, A. and A. Madabhushi, *EM-based segmentation-driven color standardization of digitized histopathology*. SPIE Medical Imaging. Vol. 8676. 2013: SPIE.

88. Mosquera-Lopez, C. and S. Agaian, *Iterative local color normalization using fuzzy image clustering*. SPIE Defense, Security, and Sensing. Vol. 8755. 2013: SPIE.

89. Doyle, S., et al., *Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer.* BMC Bioinformatics, 2012. **13**: p. 282.

90. Nguyen, K., A. Sarkar, and A.K. Jain, *Prostate cancer grading: use of graph cut and spatial arrangement of nuclei.* IEEE Trans Med Imaging, 2014. **33**(12): p. 2254-70.

91. Niazi, M.K.K., et al., *Visually meaningful histopathological features for automatic grading of prostate cancer.* IEEE journal of biomedical and health informatics, 2016. **21**(4): p. 1027-1038.

92. Sparks, R. and A. Madabhushi, *Statistical shape model for manifold regularization: Gleason grading of prostate histology.* Computer Vision and Image Understanding, 2013. **117**(9): p. 1138-1146.

93. Naik, S., et al. *Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information*. in *MIAAB workshop*. 2007. Citeseer.

94. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

95. Shwartz-Ziv, R. and N. Tishby, *Opening the black box of deep neural networks via information.* arXiv preprint arXiv:1703.00810, 2017.

Chapter 2

A version of this chapter has been submitted to Journal of Medical Imaging for publication and is currently under review: Wenchao Han, Carol Johnson, Mena Gaed, Jose A. Gomez-Lemus, Madeleine Moussa, Joseph Chin, Stephen Pautler, Glenn Bauman, and Aaron Ward, "Automatic cancer detection on digital histopathology images of mid-gland radical prostatectomy specimens."

2 Automatic cancer detection on digital histopathology images of mid-gland radical prostatectomy specimens

2.1 Introduction

Radical prostatectomy (RP), the removal of the prostate gland, is the most common treatment for organ-confined prostate cancer (PCa), which is performed on approximately 40% of prostate cancer patients annually in the United States [1]. Approximately 17%–29% of patients experience cancer recurrence after surgery, portended by serum prostate-specific antigen (PSA) relapse [2, 3]. Adjuvant and salvage therapy, including radiation therapy to the prostate bed, can be life-saving for patients with recurrence to prevent or delay mortality due to metastatic disease [3].

The surgical pathology report provides valuable information for post-surgery prognosis, recurrence risk management, and selection and guidance of adjuvant therapy. The primary and secondary tumours in the specimen are reported in terms of location, volume, and the aggressiveness of differentiation based on the Gleason grading system, which classifies the tumours into five grades based their morphological patterns, considering Gleason grade 1 – 3 as low-grade and 4 – 5 as high-grade [3].

Advances in pathology have included establishing standards for important report elements and a move towards synoptic summaries [4]. However, current clinical pathology reporting is still primarily text-based, qualitative, and subject to inter-observer variability, which leads to challenges in terms of quantitative and repeatable interpretation of lesion size, location, and spread. Methods of measuring and reporting tumour volume vary, and no generally accepted standard has been established [5]. Large inter-observer variability has been reported for identifying extra prostatic extension (EPE) (i.e. where the tumour extends outside of the prostate into the surrounding region) without anatomic landmarks [6]. In addition, assessing the differentiation degree of the tumour using Gleason score [7] (i.e. assigning a total Gleason score using the sum of Gleason grades of primary and secondary tumours) has been established for decades, but reporting the total Gleason score remains problematic. For cases with multiple nodules, reporting the overall score may underestimate the tumour aggressiveness [8].

Currently, pathology reports include or seek to incorporate accurate and detailed information to maximize clinical utility [9, 10]. Bettendorf et al. [9] proposed a hand-drawn tumour map of the prostatectomy specimen for pathology reporting which provides a way for visual estimation of tumour size and location. To quantify the hand-drawn loci on the anatomical maps in prostatectomy specimens, Eminaga et al. [10] developed an extensible markup language (XML) based document architecture. However, this method requires substantial effort from the pathologist, with potential to slow the workflow and increase fatigue. Also, the quantitative reporting based on these approaches is derived from approximate hand-drawn representations. Thus, the clinical challenges remain. Annotating each cancerous region of interest (ROI) at high resolution

(20X or higher) on whole-mount RP sections enables quantitative reporting, giving measurements of tumour size, location, and grade. This would resolve the clinical challenges mentioned above and would also benefit research studies investigating the relationship between quantitative pathologic parameters, such as tumour volume, and clinical outcomes. Moreover, PCa is challenging to detect and localize on imaging, such as multi-parametric magnetic resonance imaging. This has motivated the undertaking of imaging validation studies that use annotated prostatectomy histology as the reference standard [11, 12]. Adequate target delineation of the tumour through such studies can potentially improve disease control by allowing safe boosting of radiation dose (or targeting other ablative therapies) to corresponding areas of the prostate bed and reduction of side effects by reducing the treatment margins [13].

The time required to conduct such contouring manually precludes its use within a clinical pathology workflow and adds substantial time and expense to imaging validation studies. Therefore, there is an unmet need for a software tool that can provide accurate and fast automatic contouring of cancerous regions on whole-mount digital histopathology images of radical prostatectomy specimens.

Distinguishing cancer from non-cancer tissue on histology is challenging since their appearance are visually similar (e.g. in Figure 2.1, Gleason grade 3 (G3), small gland G4, and benign prostatic hyperplasia (BPH) (non-cancer), are visually similar). Previous studies have demonstrated the potential for computational approaches to address this problem (see the recent review article [14]). Most studies involved training and validation on a small subset (typically hundreds) of selected ROIs. In comparison, detecting cancerous regions throughout entire whole-mount slide images (WSIs) of RP

sections from different patients is more challenging for three primary reasons. First, the high resolution of the digital histopathology images leads to a large number of pixels in each WSI (e.g. a WSI of a mid-gland prostate section often includes > 4 gigapixels), requiring an efficient approach. Second, there is substantial heterogeneity of appearance in cancerous and non-cancerous prostate tissue. Non-cancerous tissue includes normal parenchyma, high-grade prostatic intraepithelial neoplasia (PIN), cancer-mimicking atrophy, and BPH, all of which having different appearances and some of which (e.g. PIN) sharing many common features with cancerous tissue. For cancerous tissue, there are different morphological patterns across different Gleason grades (e.g. different appearance for cancer samples G3 and G4, and non-cancer samples Benign, PIN, and BPH in Figure 2.1). Even within the same grade, the tissue appearance can be quite heterogeneous [3]. For instance, Gleason grade 4 has several subtypes which have different morphological patterns (e.g. small glands, large cribriform glands, mucinous glands, poorly formed glands, etc.) (Figure 2.1). Testing a system on all of the tissue throughout the WSIs thus introduces more heterogeneity into the samples, resulting in a more challenging classification problem. Third, staining variability is a substantial issue, with contributing factors such as slide preparation and batch effects [15, 16]. Although many color normalization methods have been proposed [17-19], colour normalization on all images to a template did not reduce instability of texture features and therefore has not completely resolved the problem [15]. Other challenges arise from factors including the presence of tissue marking dye due to the inking of the surfaces as part of pathology processing, artifacts (e.g. tissue folding, out-of-focus regions, tissue tearing), presence of red blood cells, and within-ROI heterogeneity (i.e. ROIs contain both cancerous and non-

cancerous tissue). Validation using all available tissue avoids bias from ROI selection and tests the system against the full variability in staining and tissue appearance.

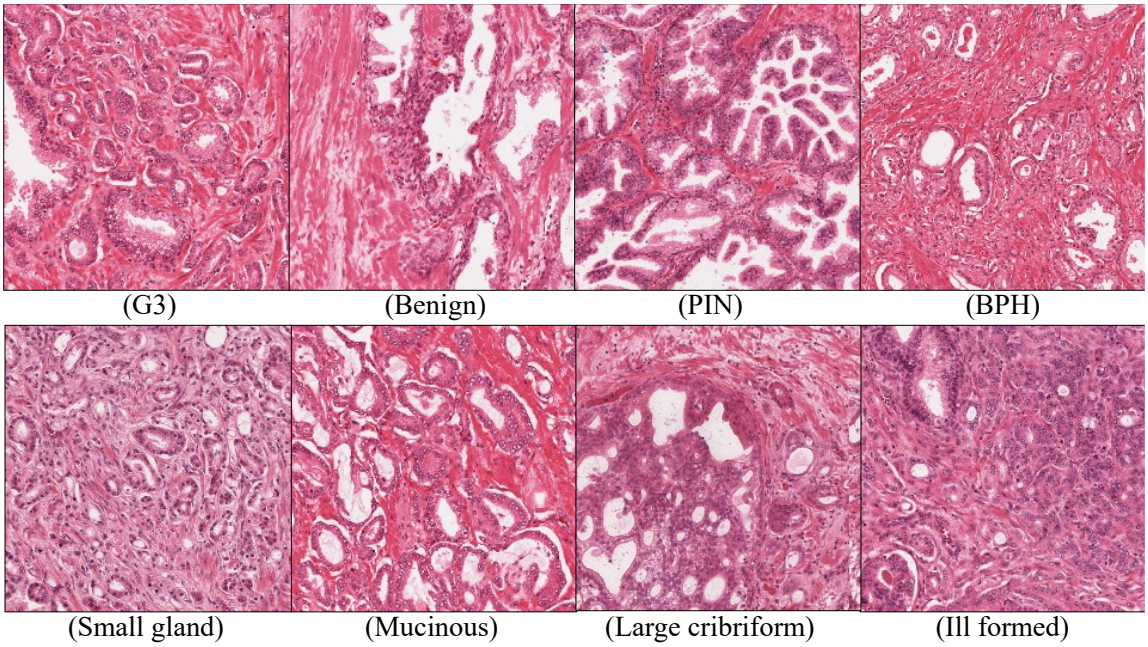


Figure 2.1: Tissue samples. Top row are cancer (G3) and non-cancer tissue samples (Benign, PIN, and BPH). Bottom row are sub-types of cancer (G4).

Chapter 2 presents a system for cancer detection and localization on WSIs of mid-gland RP sections, yielding state-of-the-art accuracies using three different methods with fast processing times. To the best of our knowledge, we have validated the system on the largest reported annotated data set, with highly detailed reference standard contours provided by an expert pathologist. The primary contributions of this chapter are:

(1) A calibration-free adaptive thresholding algorithm for fast and accurate nuclei segmentation, which yields consistent 3-class tissue component maps (TCMs) despite staining variability across WSIs.

(2) Validation of a conventional machine learning approach with 3 classifiers, and a deep learning approach using the inputs of raw images and TCMs for cancer detection on 286 WSIs of mid-gland RP sections from 68 patients (4–5 WSIs/patient). Cross-validation (CV) was performed grouping data on per-patient basis (i.e. tissues from a single patient never appeared in both the training and testing sets).

(3) Identified 14 TCM based texture features for effective cancer detection from the conventional machine learning approach, yielding better than state-of-the-art performance compared to other non-deep learning-based methods.

(4) Achieved the best overall performance across all methods using deep-learning approaches, and the use of the 3-class TCM as input reduced the sensitivity to sample size.

2.2 Related work

Substantial work has been published on the problem of prostate cancer detection on digital histology images of hematoxylin and eosin (H&E)-stained specimens [14]. Most previous research has focused on the classification of cancerous vs. non-cancerous using pre-selected ROIs, with a few previous studies focused on cancer identification and localization on whole-mount tissues (biopsy and RP).

Studies [20-28] performed cancer vs. non-cancer classification using pre-selected ROIs. Their overall positive results are valuable in that they point to the potential utility of the methods employed. However, the total processed tissue amount in each study was less than 4 cm^2, which is about half of the size of a single WSI of a mid-gland tissue section. Considering the heterogeneity of the prostate cancer tissue and staining

variability among WSIs, the limited extent of the processed tissue points to the need for a study with more comprehensive validation throughout entire WSIs. Also, these studies generally did not prioritize computational cost and the scalability to larger tissue samples as would be encountered in clinical practice.

Several proposed approaches have demonstrated the potential for detection and localization of cancerous regions on WSIs of biopsy tissues. Doyle et al. [29] used a multiresolution approach to detect PCa with a boosted Bayesian classifier, reported areas under the receiver operating characteristic curve (AUCs) of 0.84, 0.83, and 0.76 at the lowest, intermediate, and highest resolution levels respectively on 100 biopsy WSIs from 58 patients with average processing time of approximately 3 minutes per 1000 ×1000 pixel region. Litjens et al. [30] used a super-pixel based approach and reported an AUC of 0.96 at the per-slide level with a sensitivity of 1.0 and specificity 0.4, and an AUC of 0.92 at the super-pixel patch level using 10-fold CV (data were stratified on a per-patient basis) using 204 WSIs of biopsy tissues from 163 patients. Approximately 4 minutes of processing time per WSI was reported. Their more recent work [31] used a convolutional neural network and reported an AUC of 0.99 (90th percentile) and 0.98 (median percentile). The processing time was not reported. Overall, it is not clear that these systems would scale up to the data sizes involved in RP specimens; for instance, a computation time of 4 minutes per WSI of biopsy tissue [30] would result in a requirement of approximately 17 days of computation per mid-gland RP WSI.

Several studies reported system designs for detecting PCa on WSIs of prostate tissue sections. Monaco et al. [32] and Rashid et al. [33] classified each individual gland as malignant or benign by extracting gland features. Monaco et al. reported a sensitivity

and specificity of 0.87 and 0.90 respectively using 40 RP tissue sections from 20 patients at low resolution (8μm/pixel) with average processing time of 2.75 minutes per whole-mount tissue section. Rashid et al. reported a sensitivity of 90%, a specificity of 93%, and an accuracy of 93%, validating on 20 WSIs from 11 patients with a training data set of 50 WSIs from 19 patients without reporting system processing time. Both works reported a limitation that the system was unable to detect poorly differentiated cancer due to the dependence of the method on gland classification. This is an issue for Gleason grade 5 (G5) cancer, in which the glands are disrupted. In their data set, no G5-inclusive tissue samples (i.e. pure G5 and mixed grades including G5, such as G4+5) were used. However, the presence of G5 cancer is highly negatively prognostic [34] and therefore it is critical to detect G5 cancer.

DiFranco et al. [35] used a tile-based approach, classifying each 512×512 pixel tile of the WSI from RP specimens and reported an AUC of 0.95 using 15 WSIs from 14 different patients. Texture features were calculated at different colour channels of RGB and CIE L*a*b colour space. The experiments conducted in the study used a 16-core server without reporting computational time, and the need for parallel computing was described.

Nguyen et al. [36] reported a false positive rate of 6% with 78% sensitivity using 6 images (approximate size of 4000×7000 pixels) for training and 11 WSIs (approximate size of 5000×23000 pixels) for testing, incorporating nucleus-related cytology features with texture features without reporting the processing time.

Nir et al. [37] used multiple pathologists' annotations with hand-crafted features and reported an AUC of 0.85 with an accuracy of 90.5%, a sensitivity of 91.5%, and a specificity of 85.2% at the optimal operating point. They tested on 333 tissue microarray (TMA) cores sampled from 231 RP specimens for classifying TMA patches as cancerous vs. non-cancerous. The validation was conducted using leave-one-patient-out (LOPO) CV with processing time of approximately 14 hours for feature extraction for 333 TMAs, and approximately 4.5 hours for 231 LOPO simulations for training and classification using parallel computing via two 12-core computers. They reported an overall AUC of 0.75 when testing their system, trained using all of the TMAs, on an external dataset of 230 WSIs of RP sections from 56 patients using all of the cancerous tissues, and 10% of the non-cancerous tissue by random sampling.

To the best of our knowledge, no system for prostate cancer detection has been reported and validated for performance throughout all tissues on RP WSIs including all clinically relevant grade groups, with practical processing time.

2.3 Methods

Figure 2.2 is a block diagram depicting the training of conventional and deep learning-based classifiers for automatic cancer detection on WSIs of RP sections, building on our previously reported prototype system [38]. In training, cancerous and non-cancerous ROIs were determined using gold-standard histopathology annotations. We computed a 3-class TCM for each ROI using our proposed segmentation method. The three machine learning approaches used in our study were as follows. (1) We computed 14 selected first- and second-order statistical features [39, 40] from TCMs and trained three classifiers (a Fisher linear discriminant classifier [FisherC], a Logistic linear

classifier [LoglC], and a support vector machine classifier [SVM]). (2) Pre-trained AlexNet [41] (pre-trained on the ImageNet database [42]), was fine-tuned using the TCMs. (3) Pre-trained AlexNet was fine-tuned using raw image ROIs without any image processing. During testing, the trained system classified all of the ROIs covering each WSI as cancerous or non-cancerous. We validated the classification results against gold standard histopathology annotations using CV, with data grouped on a per-patient basis, ensuring that samples from the same patient never appeared in both the training and testing sets in any fold. We performed the validation using each of the trained systems, and the results were collected for analysis and comparison. Our implementation used Matlab 2018a (The Mathworks, Natick, MA), OpenCV 3.1 for SVM implementation, and PRtools 5.0 (Delft Pattern Recognition Research, Delft, The Netherlands) for the implementation of FisherC and LoglC.

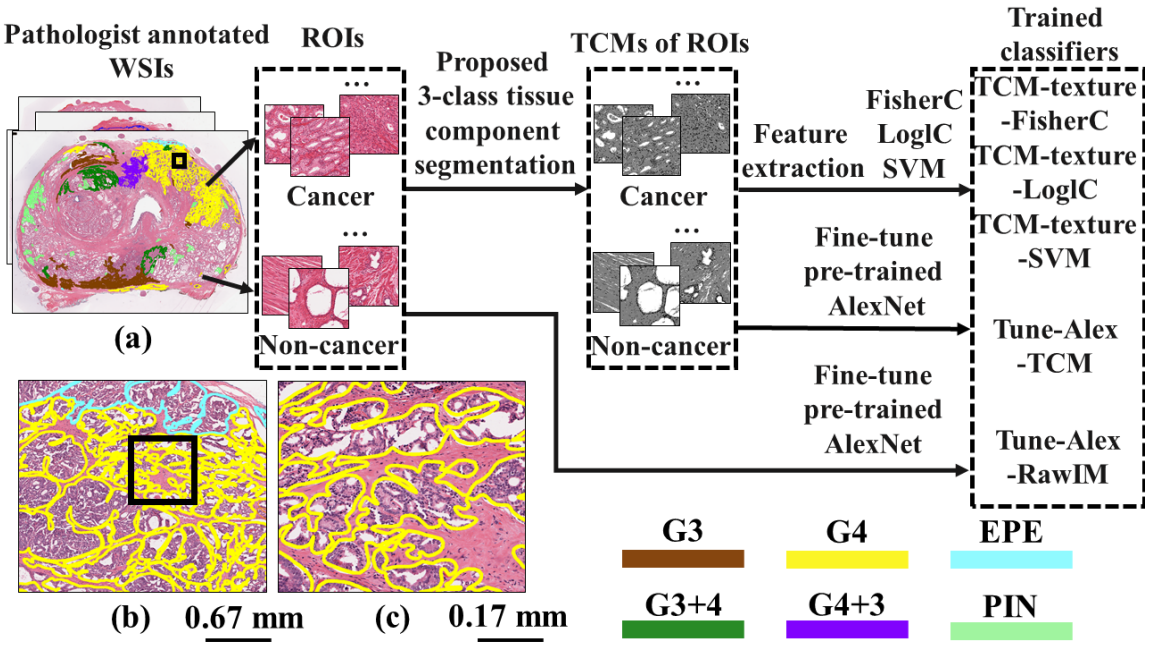


Figure 2.2: Method overview for system training using 3 different machine learning methods. (a) WSIs with our expert annotations. Different coloured annotations represent different types of tissue based on the Gleason grading system. (b) and (c) are zoomed views from the black square regions in (a) and (b) respectively.

2.4 Data

2.4.1 Manual annotation

This study was approved by our institutional Human Subjects Research Ethics Board with informed consent of all patients. We obtained 299 WSIs of H&E-stained, 4 μm thick, paraffin-embedded mid-gland tissue sections from 71 radical prostatectomy specimens from patients with biopsy-confirmed prostate cancer (clinical stage T1 or T2). All tissue sections were prepared and scanned in our hospital pathology laboratory following the same protocol [26]. Two different types of scanners were used: an Aperio ScanScope GL (Leica Biosystems, Wetzlar, Germany) for sections from 46 patients and

an Aperio ScanScopeAT Turbo (Leica Biosystems, Wetzlar, Germany) for sections from the other 25 patients. There is no difference in scanning specifications between the two models. Sections were scanned at 20X (0.5μm/pixel) in bigtiff pyramid format without compression. The resulting 24-bit RGB colour images have a pixel size of 0.5μm/pixel.

Each WSI was contoured and graded by a trained physician (Gaed) at 20× magnification using different colours for different Gleason grades, and verified by one of two genitourinary pathologists (Moussa or Gomez). Contouring was performed using a Cintiq 12WX pen-enabled display (Wacom Co. Ltd., Saitama, Japan) with the ScanScope ImageScope v11.0.2.725 image viewing software (Aperio Technologies, Vista, CA, USA). Contouring was conducted at high precision (Figure 2.2 (b, c)), which takes about 70 hours per case. Where Gleason grades were intermingled to a degree where they could not be readily separated, foci were given a grade such as 4+3, indicating that the majority of the focus contained grade 4 cancer, and the remainder contained grade 3 cancer.

2.4.2 Ground truth ROI labeling

Each WSI was separated into a set of square 960×960 pixel ROIs. ROIs containing at least 50% cancerous tissue according to the manual pathology annotations were considered cancerous; all other ROIs were considered non-cancerous.

2.4.3 Data separation for system tuning and feature selection

We performed classifier hyper-parameter tuning and feature selection on a separate “tuning data set” comprising 13 WSIs from 3 different patients. We did not use the patients in the tuning data set for CV; we used only the 68 remaining patients for CV. The tuning and CV data sets each have a mixture of WSIs from both scanners.

2.5 Tissue component mapping

The purpose of tissue staining is to assist in identifying different types of tissue components which have semantic meaning to the pathologist for identifying abnormalities. We developed an algorithm that assigns a label to each image pixel to generate a TCM for further analysis. We labeled each pixel as one of the three tissue components: nuclei, lumen, and stroma/other tissue. The main steps of our TCM generation algorithm are: (1) segmentation of nuclei using colour deconvolution and a proposed adaptive thresholding method, (2) segmenting luminal areas by global thresholding in the red-green-blue (RGB) colour space, and (3) designating other pixels as stroma/other tissue. These steps are described in more detail as follows.

2.5.1 Nucleus mapping

We segmented cell nuclei by adaptive thresholding of the hematoxylin channel after colour deconvolution [43] to compensate for staining variability that observed across different WSIs in our dataset. We then used morphological operations to reduce red blood cells (RBCs) pixels that were falsely labeled as nuclei, by relabeling them as stroma/other tissue. The details of these steps are described below.

Colour deconvolution: We used a colour deconvolution algorithm [43] to separate the H&E stains into three image channels corresponding to the hematoxylin stain, eosin stain, and the background.

Following the Beer-Lambert law [44], y=CM; i.e. the optical density y detected for a particular pixel is linear with respect to the stain amount C, where M is the optical density matrix for the colour stains. In our case, the colour stains are hematoxylin, eosin,

and background. The stain-specific values for the optical density (matrix M) in each of the three channels can be determined by measuring relative absorption for red, green, and blue on slides stained with a single stain [43]. It is straightforward to see that $C = Dy$, where $D = M^{-1}$ is the colour deconvolution matrix.

We used the standard deconvolution matrix used by Ruifrok and Johnston [43] and applied this algorithm to each ROI independently. This separated each ROI into three grey-level images representing the amount of hematoxylin (e.g. Figure 2.3 (b), darker region corresponding to larger amount of hematoxylin stain), eosin, and background respectively. Since most substances within nuclei are basophilic, they bind to hematoxylin. We therefore used the hematoxylin channel to segment nuclei.

Cell nucleus segmentation using adaptive thresholding: Staining variability results in variations in hematoxylin channel intensity across different WSIs (see the middle images for the two sample ROIs in Figure 2.4 (a)). The left case has much more darkly hematoxylin-stained stroma/fibromuscular tissue than the right, where the nuclei are more prominent. This makes global thresholding inapplicable for nucleus segmentation. We therefore propose an adaptive thresholding method. The algorithm computes an optimal threshold for each WSI. This threshold is used as a global threshold for the WSI to segment the nuclei pixels from the grey-level images of hematoxylin channel of each ROI after colour deconvolution.

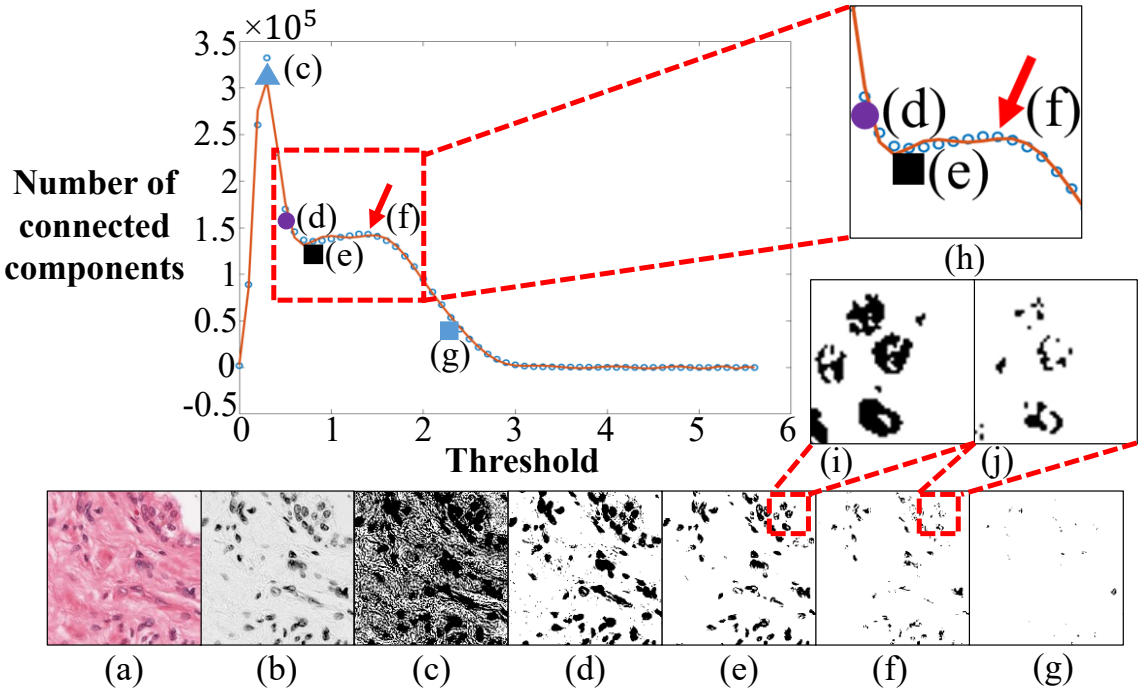


Figure 2.3: Plot of number of connected components and the binarized hematoxylin channel at the corresponding thresholds. (a) Sample ROI. (b) Grey-level image representation of the hematoxylin channel after colour deconvolution. (c), (d), (e), (f), and (g) are binary maps after thresholding using the thresholds where the blue triangle, purple circle, black square, red arrow, and blue square labeled in the plot respectively. The red square highlighted region in a zoomed in view shows in (h). (i) and (j) are images of zoomed in view highlighted by the red squares in (e) and (f) respectively

Our proposed algorithm is based on the observation that as the threshold on the hematoxylin channel increases, causing more hematoxylin-stained tissue to be excluded in the thresholded image (Figure 2.3 (c–g)), initially the components of background tissue are excluded as a sharp decreasing number of connected components (Figure 2.3 (c–d)), resulting in a nuclei map with some background tissue. With further increase of the threshold, background tissue were excluded to form a nuclei-only map (Figure 2.3 (e)).

At this point of inflection, the threshold can separate nuclei from other tissue components (Figure 2.3 (e)). With further increasing of the threshold, the number of connected components increase slightly (shown qualitatively in Figure 2.3 (e–f) and quantitatively in Figure 2.3 (h) between the black square and the red arrow) because pixels from the same full nucleus disappear to separate the nucleus to form multiple independent connected components (Figure 2.3 (e, f, i, j)).

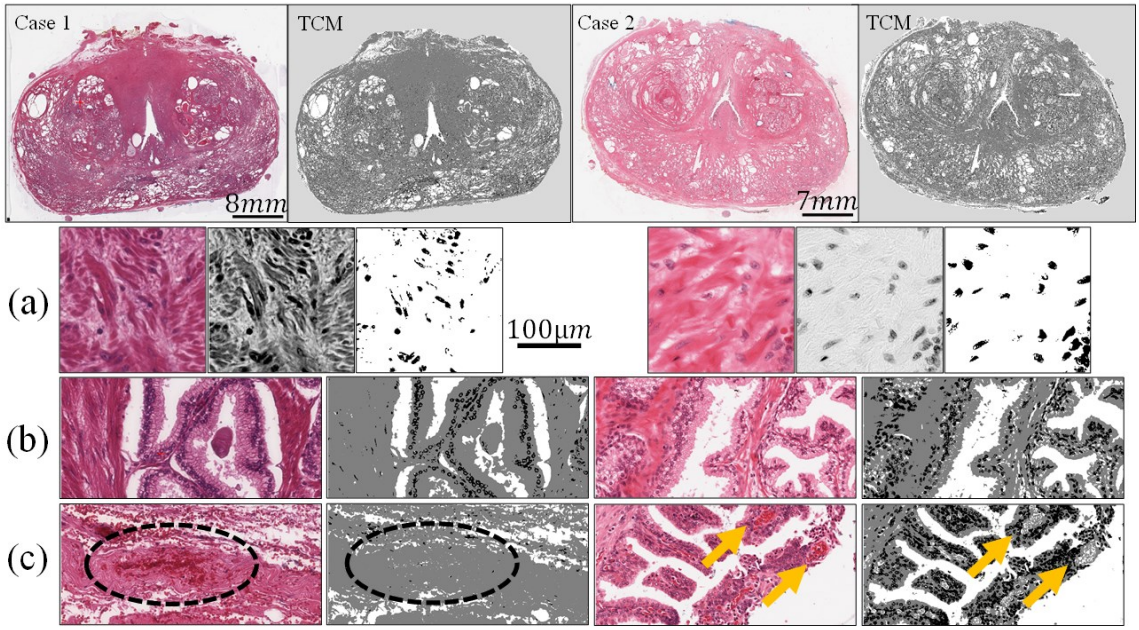


Figure 2.4: Tissue component segmentation. Top row: Two cases of H&E stained WSIs with their TCMs to the right. (a) Left: ROI samples from each of the cases, middle: grey-level hematoxylin channel images, right: nuclei map after adaptive thresholding. (b) Left: ROI samples from each of the cases, right: computed TCM using our segmentation method. (c) Left: ROI samples from the two cases with RBCs included, right: TCM computed using our segmentation method. RBCs are circled in the first case and pointed by yellow arrows in the second case.

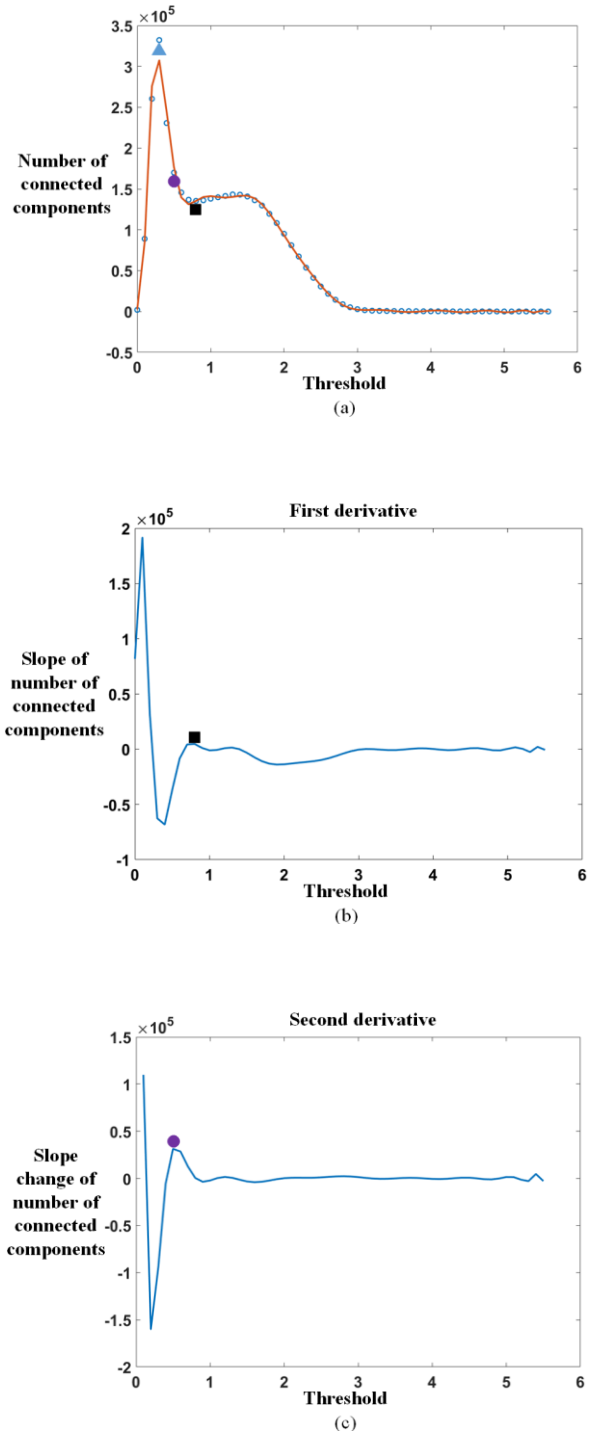


Figure 2.5: Plots of (a) number of connected components and amount of hematoxylin stain with fitted curve in red, (b) first derivative of curve in (a), (c) second derivative of curve in (a).

To detect this inflection point, we first fit a high-order polynomial curve to the data depicting the number of connected components as a function of threshold. The hyper-parameter of curve order was tuned manually to 20 using our tuning data set, with two objectives in mind: (1) making the polynomial order as low as possible to avoid noise generated from oscillation based on the Runge phenomenon [45], and (2) minimizing the squared error between the curve and the original points. The resulting continuous and differential plane curve was fit using least squares approximation, and the coefficients were calculated using the Vandermonde method [46] with the chosen curve order.

Using the fit curve, we computed the desired inflection point in three steps. First, we computed the threshold $\tilde{x}$ giving the largest number of connected components (blue triangle in Figure 2.3 and Figure 2.5 (a)) as denoted,

$$\tilde{x} = \arg\max_{x \in X} F(x) \tag{2.1}$$

where x is a threshold on the hematoxylin channel, X is the domain of thresholds on the hematoxylin channel and F is the fit polynomial curve. Second, we computed the threshold $\ddot{x} > \tilde{x}$ corresponding to the most rapid decrease in connected components (purple circle in Figure 2.3 and Figure 2.5 (a, c)) as denoted,

$$\ddot{x} = \arg\max_{x \in X,\ x > \tilde{x}} F''(x) \tag{2.2}$$

Third, the threshold $x_T \in X$ is defined as the closest rising inflection point to $\ddot{x}$ (black square in Figure 2.3 and Figure 2.5 (a, b)):

$$x_T = \underset{x \in X_T}{\operatorname{argmin}}\, x - \ddot{x} \tag{2.3}$$

$$X_T = \underset{x_0 \in (x-\delta,\ x+\delta),\ x>\ddot{x}}{\arg} F'(x) > F'(x_0) \tag{2.4}$$

where $(x - \delta,\ x + \delta)$ is the local neighborhood of x. X_T is the set of isolated local maxima of the first derivative (therefore the set of rising inflection points) of curve F such that $x_T > \ddot{x}$.

This threshold was found for each WSI via a cumulative assessment of 2,000 randomly-selected $120\mu m \times 120\mu m$ samples lying within the prostate (i.e. avoiding clear slide areas) and not containing tissue marking dye (i.e. avoiding areas of artefact).

RBC removal: Since RBCs (Figure 2.6 (a, f)) also stain with hematoxylin, the adaptive thresholding process erroneously labels them as nuclei (Figure 2.6 (b)). However, RBCs have higher saturation than nuclei in a red-pink hue, allowing us to distinguish them from nuclei. We selected and applied hue-saturation-intensity RBC thresholds (hue ≥ 0.95/1, saturation ≥ 0.72/1, and intensity ≥ 0.6/1) based on a cumulative histogram from 100 $40\mu m \times 40\mu m$ RBC ROIs selected from our tuning data set (see a sample thresholding result in Figure 2.6 (c)). We then applied morphological dilation with a disk-shaped structuring element of radius = $4\mu m$ (approximate radius of human red blood cells) (Figure 2.6 (f)). This resulted in an RBC mask (Figure 2.6 (e)) that was subtracted from the nucleus map (Figure 2.6 (b)) to eliminate falsely detected RBCs (Figure 2.6 (d)).

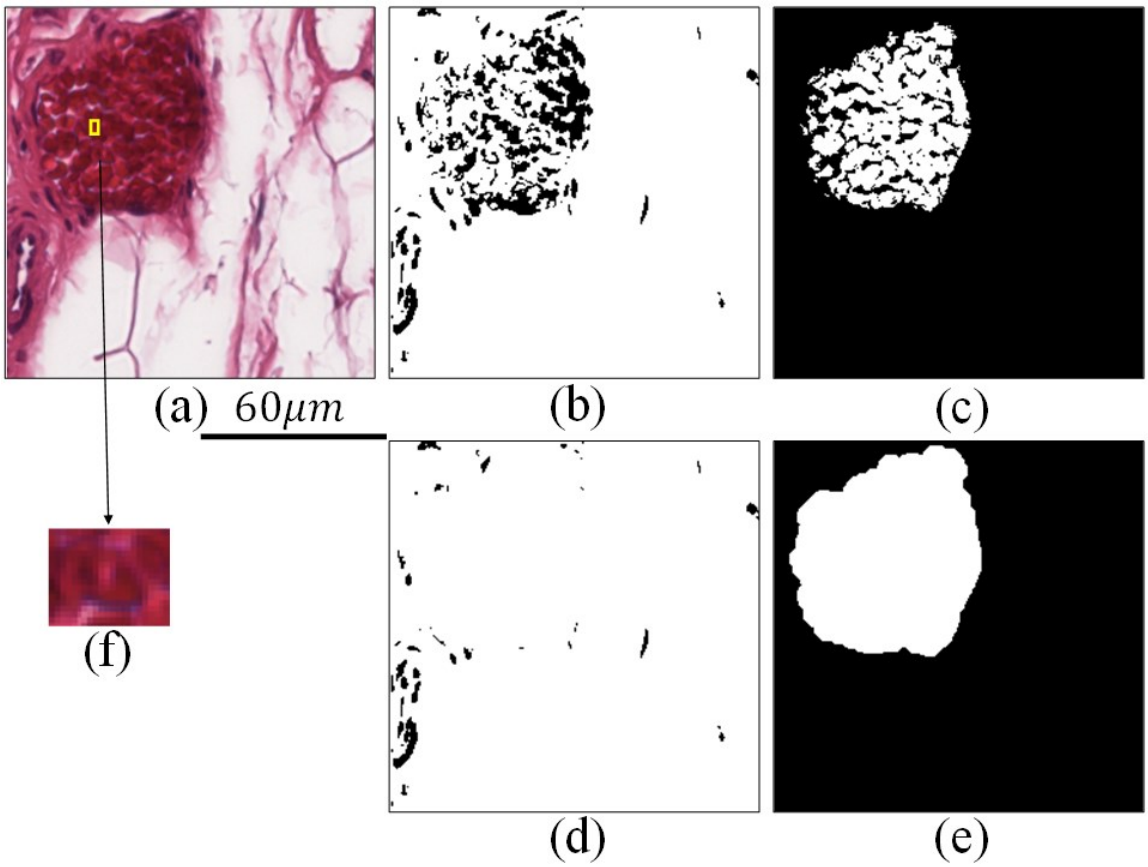


Figure 2.6: Red blood cell (RBC) removal for an example ROI (a). (b) and (d) are nuclei maps before and after RBC removal respectively. (c) is a binary mask covering the RBCs, generated by thresholding in HSV colour space. (e) is a binary mask created from (c) after morphological operation with a disk shaped structuring element of radius = 4μm (approximate radius of human red blood cell (f)).

2.5.2 Lumen and stroma/other tissue component segmentation

Luminal regions are consistently nearly white on each WSI. We used global thresholding to segment luminal pixels with threshold values of red ≥ 0.86/1, green ≥ 0.71/1, and blue ≥ 0.82/1 (threshold values were chosen using a cumulative histogram

calculated from luminal ROIs sampled from the tuning data set). All pixels that were not labeled as nuclei or lumen were labeled as stroma/other.

2.5.3 Tuning ROI size and down-sampling ratio

The ROI size used in the experiment affects the resolution of the resulting cancer map, and the down-sampling ratio of the TCM affects computation time. We selected an ROI size of 480μm×480μm (960×960pixel) and nearest-neighbor down-sampling ratio of 0.25 by experimentation with our tuning data set. We tested ROI sizes from 120μm×120μm to 720μm×720μm in 120μm steps. We tested down-sampling ratios of 0.25 to 1 in increments of 0.25. The selection of those two parameters was manually performed based on evaluating the performance (area under the receiver operating characteristic curve) for cancer detection (using FisherC) in leave-one-patient-out cross-validation on the tuning data set.

2.6 Feature extraction and selection

We calculated 24 first-order and 132 second-order statistical features [39, 40] from the TCM of each ROI, giving a total of 156 features. The second-order statistical features were based on the grey-level co-occurrence matrix (GLCM) [39]and grey-level run length matrix (GLRLM) [40]. GLCMs and GLRLMs were calculated using neighbors in four directions [(0,1) denoted direction 1 in Table 2.1, (-1,1) denoted 2, (-1,0) denoted 3, and (-1,-1) denoted 4] without aggregation over the directions. In total, we calculated 156 features: (22 GLCM + 11 GLRLM) × 4 directions + 24 first-order features =156.

We selected the 14 top ranked features using backward feature selection via ranking the AUCs from leave-one-patient-out cross-validation (LOPO CV) of a Fisher

linear classifier with the TCMs on the tuning dataset. The number of features was chosen by iterative experiments from 1 to 156 using feature selection as described. The chosen texture features are listed in Table 2.1.

Table 2.1: Selected features used in cross validation

Mean gradient value
GLRLM Short Run Low Gray Level Emphasis-1
GLRLM Short Run Low Gray Level Emphasis-3
GLRLM Short Run Low Gray Level Emphasis-4
GLCM Energy-1
GLCM Energy-2
GLCM Information Measure of Correlation-1
GLCM Information Measure of Correlation-2
GLCM Inverse Difference Moment -2
GLCM Inverse Difference Moment -3
GLCM Cluster shade-3
GLCM Correlation-1
GLCM Entropy-2
GLRLM Short Run Emphasis-3

2.7 Cancer detection using machine learning

We classified each ROI as cancerous vs. non-cancerous using the calculated features. We performed supervised machine learning on the TCMs using (1) FisherC, (2) LoglC, and (3) SVM (NU-SVC with a radial basis function kernel, parameters tuned as cost = 12.5, gamma = 0.50625 using our tuning dataset). Each of these classifiers is henceforth denoted as: (1) TCM-Texture-FisherC, (2) TCM-Texture-LoglC, and (3) TCM-Texture-SVM, respectively.

We also used transfer learning by fine-tuning pre-trained AlexNet with our TCMs and raw image ROIs, which denoted as: (4) Tune-AlexNet-TCM, and (5) Tune-AlexNet-

RawIM. AlexNet is a convolutional neural network. It consists of 5 convolutional layers followed by 3 fully connected layers with one classification layer of 1000 outputs. The model demonstrated a winning performance in the ILSVRC-2012 challenge. The model was trained by 1.2 million training images from ImageNet [42], which is a natural image data base [41]. The idea is to tune the pre-trained model, which is trained by natural images, with our histopathology image ROIs to train a system for classifying cancer vs. non-cancer ROIs. The methodological details of using transfer learning are described in the following paragraphs.

We used TCMs as input images to fine-tune the pre-trained AlexNet. The TCMs were converted into RGB colour images, in which red, green, and blue represent nuclei, stroma/other, and lumen respectively. All the ROIs of size of 240×240×3 were resized to 227×227×3 (to conform to the necessary input size for AlexNet) using bilinear interpolation. For comparison, the same method as that used for Tune-AlexNet-TCM was repeated using the "raw" unmodified H&E images instead of TCMs.

We replaced the final fully connected layer of AlexNet with a fully connected layer, which has a 2-way output followed by the 2-way softmax algorithm with a 2-class label output (cancerous vs. non-cancerous). We calculated the loss function using cross-entropy. The weights and biases of the replaced layers were initialized with random numbers. We set the initial learning rate $\alpha = 0.0001$ for all the other layers, and $\alpha = 0.002$ for the output layer to make the weights and biases from other layers almost unchanged while those from the output layer learn faster. We used the adaptive moment estimation ('Adam') optimizer [47] for gradient descent. The following hyper-parameters were set as: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$ [47]. Other training parameters were set as: mini-

batch size = 200, maximum epoch = 10, which (including α) were chosen by using our tuning dataset.

2.8 Experimental design and evaluation methods

2.8.1 Cross validation

We performed LOPO, 5-fold and 2-fold CV for each classifier to classify each ROI as cancerous vs. non-cancerous covering each WSI. In each CV, data were grouped on a per-patient basis, such that no same-patient samples were used in both the training and testing sets. During training, the positive (cancerous) and negative (non-cancerous) samples were balanced by random subsampling of the negative samples. We performed testing on all ROIs covering each WSI in our 68-patient set (i.e. all tissue on all slides was classified; we did not use only selected ROIs in our experiments).

We calculated the error metrics (using a fixed operating point at the confidence level of 0.5) of error rate, false positive rate (FPR), false negative rate (FNR), and AUC by validating against our expert manual annotations. The overall performance was measured by averaging each of the error metrics across all the patients. The sample sizes for each tissue type are shown in Table 2.2. We also calculated the error rates (FNRs for cancerous tissue types; FPRs for non-cancerous tissue types) for each tissue type using LOPO CV.

We did statistical tests to compare the AUCs obtained from the different methods. We grouped AUC values into three groups: group 1 (TCM-Texture-Fisher, TCM-Texture-LoglC, and TCM-Texture-SVM); group 2 (Tune-AlexNet-TCM); group 3 (Tune-AlexNet-RawIM). We tested the AUCs for normality using the Shapiro-Wilk test.

We compared the AUCs for the three groups (i.e. group 1 vs. group 2 vs. group 3) using the Kruskal-Wallis test. We then compared each pair of groups using the Wilcoxon rank sum test.

Table 2.2: Sample size for each tissue type (480 mm × 480 mm)

Cancerous tissue types									Non-cancerous tissue types		
G3	G4	G5	EPE	G3+4	G4+3	G4+5	G5+4	G5+3	Atrophy	PIN	Healthy
14718	3949	37	272	6008	3839	727	8216	16	5433	26449	1178814

2.8.2 Training sample size experiment

We conducted an experiment to investigate how the number of patients in the training set influences the system performance for each machine learning approach. We randomly selected 34 patients as the testing set, then iteratively trained the classifiers using training set sizes ranging from 1 patient to 33 patients. At each iteration, we trained all of the classifiers described in Sec. 2.7, tested the trained systems on the entire 34-patient testing set, and computed error metrics as described in Sec. 2.8.1.

2.9 Results

2.9.1 Tissue component segmentation

The average adaptive threshold on the hematoxylin channel for WSIs within each patient ranged from 0.5 to 1.1. The standard deviation ranged from 0 to 0.4, with 42% of patients having a standard deviation ≥ 0.1. The average ± standard deviation threshold for all patients was 0.7±0.2.

Qualitative results are shown in Figure 2.4, where staining variability was observed for the two WSIs shown (Figure 2.4 (a)). The hematoxylin channel thresholds

are different for the samples after colour deconvolution (Figure 2.4 (a), middle image for each case). The nuclei were segmented by our adaptive thresholding method to generate nuclei maps (Figure 2.4 (a), right image for each case).

Figure 2.4 illustrates TCMs for the two WSIs right beside each case. Note the consistency of the maps despite the observed staining variability. Figure 2.4 (b) and (c) show zoomed views for samples taken from the two cases, and their TCMs after segmentation. Also note that the confounding red blood cells in the oval in Figure 2.4 (c) left and indicated by the yellow arrows in Figure 2.4 (c) right were correctly classified into the "stroma/other" category and not as nuclei.

Table 2.3: Error metrics for cancer vs. non-cancer classification from each cross validation

	Error rate (%)	FNR (%)	FPR (%)	AUC [0, 1]
Leave-one-patient-out cross-validation (LOPO CV)				
TCM-Texture-FisherC	13.7±6.7	12.6±16.9	13.6±7.1	0.94±0.05
TCM-Texture- LoglC	12.3±6.0	12.4±15.6	12.2±6.4	0.95±0.05
TCM-Texture-SVM	8.5±4.6	13.1±13.7	8.2±4.6	0.96±0.04
Tune-AlexNet-TCM	6.2±4.0	10.7±12.6	6.0±4.2	0.98±0.03
Tune-AlexNet-RawIM	5.5±6.9	8.5±13.8	5.2±6.7	0.98±0.02
5-fold cross validation (5-fold CV)				
TCM-Texture-FisherC	13.7±6.4	13.0±17.4	13.7±6.9	0.94±0.05
TCM-Texture- LoglC	12.1±5.6	12.7±15.8	12.1±6.0	0.95±0.05
TCM-Texture-SVM	8.4±4.4	13.6±13.7	8.2±4.5	0.96±0.05
Tune-AlexNet-TCM	7.4±5.4	10.5±11.5	7.1±5.3	0.97±0.02
Tune-AlexNet-RawIM	4.8±4.5	8.8±11.8	4.3±3.8	0.98±0.02
2-fold cross validation (2-fold CV)				
TCM-Texture-FisherC	13.7±6.1	14.0±17.7	13.5±6.5	0.94±0.05
TCM-Texture- LoglC	12.0±5.6	14.7±15.8	11.6±6.0	0.94±0.05
TCM-Texture-SVM	8.5±5.0	16.6±17.5	7.8±4.2	0.94±0.06
Tune-AlexNet-TCM	7.7±4.3	8.1±10.0	7.8±4.5	0.97±0.03
Tune-AlexNet-RawIM	9.0±7.6	6.1±10.3	9.0±7.9	0.98±0.02

2.9.2 Cancer vs. non-cancer classification

The quantitative results for cancer vs. non-cancer classification from our CV experiments are reported in Table 2.3. For LOPO CV experiments, from the normality test, AUCs were normally distributed. The Wilcoxon rank sum test results for the two groups are shown in Table 2.4. For AUCs, group 1 and 2, group 1 and 3, and group 2 and 3 were significantly different. The confusion matrix was calculated, as shown in Table 2.5, using the fixed operating point corresponding to the confidence level of 0.5. The calculated error rates for each tissue type (i.e. the FNRs for cancerous tissue types, and the FPRs for the non-cancerous tissue types) are shown in Figure 2.7.

Figure 2.8 shows the quantitative results from the training sample size experiment.

The qualitative results for the example cases are shown in Figure 2.9. It illustrates the capability of our system to map cancer throughout entire WSIs. The upper and lower cases in Figure 2.9 depict average and below average performance of the methods (see error metrics below each map in Figure 2.9 in comparison to Table 2.3) for LOPO CV using each of three classifiers.

Table 2.4: Wilcoxon rank sum test results for each of the two groups

Testing groups			p values
AUC			
Group 1	Group 2		**<0.0001**
Group 1		**Group 3**	**<0.0001**
	Group 2	**Group 3**	**<0.006**

p values that are significant ($p<0.05$) are bolded in the table. Groups that have better performance are bolded in the table.

Table 2.5: Confusion matrix for each method from leave-one-patient-out cross-validation; each sample is a 480 μm × 480 μm ROI.

TCM-Texture-FisherC		Pathologist annotation	
		Cancerous	Non-cancerous
System predicted label	Cancerous	32134	162974
	Non-cancerous	5647	1047722

TCM-Texture-LoglC		Pathologist annotation	
		Cancerous	Non-cancerous
System predicted label	Cancerous	31385	145563
	Non-cancerous	6396	1065133

TCM-Texture-SVM		Pathologist annotation	
		Cancerous	Non-cancerous
System predicted label	Cancerous	30260	99523
	Non-cancerous	7521	1111173

Tune-AlexNet-TCM		Pathologist annotation	
		Cancerous	Non-cancerous
System predicted label	Cancerous	32092	70530
	Non-cancerous	5689	1140166

Tune-AlexNet-RawIM		Pathologist annotation	
		Cancerous	Non-cancerous
System predicted label	Cancerous	30537	66983
	Non-cancerous	7244	1143713

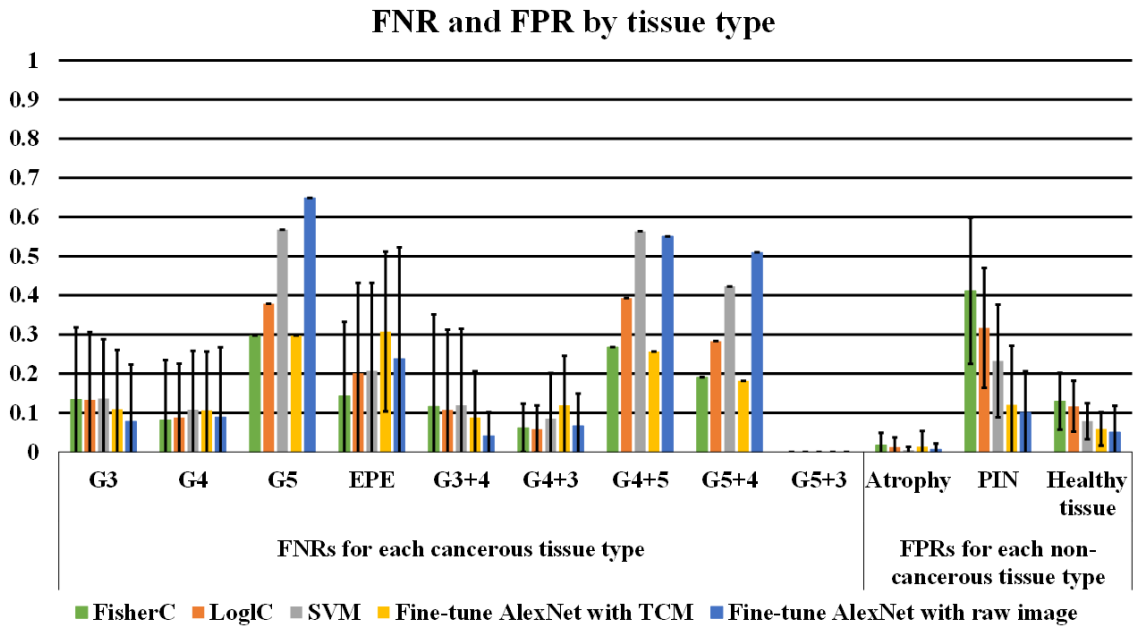


Figure 2.7: Mean ± standard deviation of FNR for each cancerous tissue type; mean ± standard of FPR for each non-cancerous tissue type (atrophy, PIN, and healthy tissue) from the LOPO CV.

Our implementation is not optimized for speed, and consists primarily of Matlab code. The non-deep learning methods require approximately 45 minutes to map cancer throughout an entire WSI using an Intel i5 workstation @ 3.10GHz- with 24 GB of random access memory. For the deep learning methods, the Tune-AlexNet-TCM method requires 15 minutes, and the Tune-AlexNet-RawIM method requires 3 minutes.

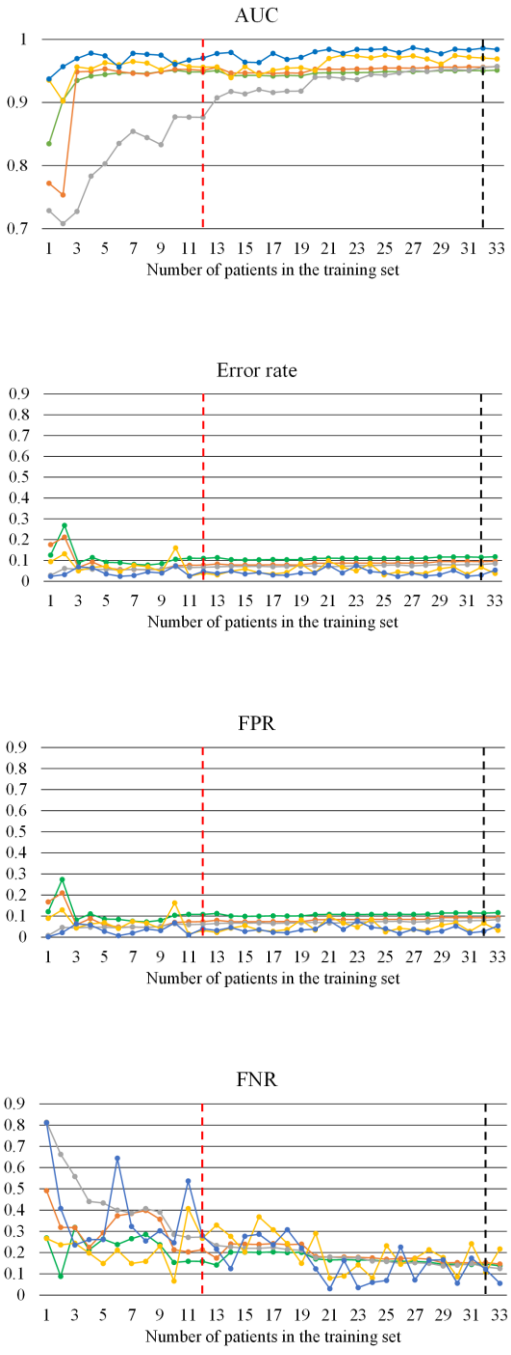


Figure 2.8: Plots for error metrics at each training patient number for cancer detection. Green: TCM-Texture-FisherC. Orange: TCM-Texture-LoglC. Grey: TCM-Texture-SVM. Yellow: Tune-AlexNet-TCM. Blue: Tune-AlexNet-RawIM. Red and black dashed lines: reference points using 12 and 32 patients for training respectively.

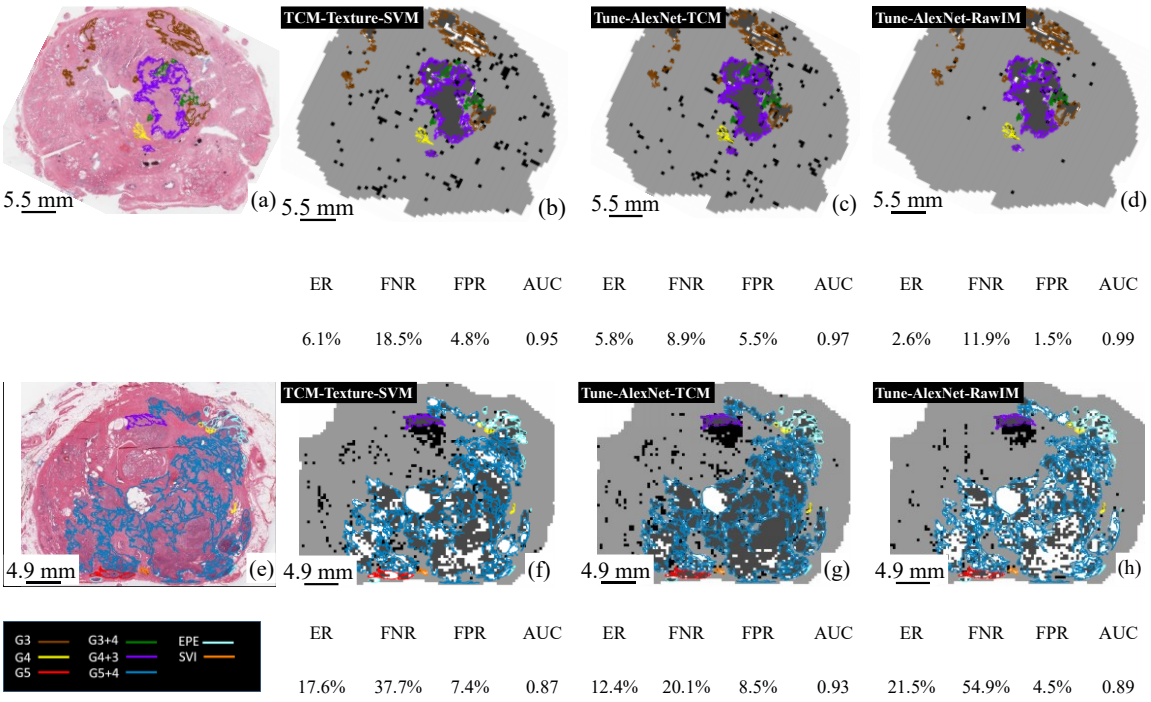


Figure 2.9: Cancer maps for two example whole slide images (WSIs). (a) and (e) are example WSIs; (b) and (f) are cancer maps from TCM-Texture-SVM; (c) and (g) are cancer maps from Tune-AlexNet-TCM; (d) and (h) are cancer maps from Tune-AlexNet-RawIM. Colour contours in each image are the pathologist's annotations. The error metrics are below each cancer map. Labels in the cancer maps are: dark grey – true positives, light grey – true negatives, black – false positives, white – false negatives. ER: Error rate.

2.10 Discussion

In this chapter, we proposed an approach to tissue component mapping that finds the loci of nuclei, luminal regions, and other tissue components (including stroma) based on our adaptive thresholding algorithm which compensates for staining variability across the WSIs. This algorithm is amenable to fast implementation and yields consistent TCMs supporting cancer detection using machine learning algorithms. We validated conventional and deep learning-based approaches for classifying $480\mu m \times 480\mu m$ ROIs as cancer or non-cancer throughout digitized mid-gland WSIs of RP sections. We did not subsample the tissue available in our data set; we cross-validated using all of the tissue on every slide, ensuring that tissues from the same patient never appeared in both the training and testing sets. For CV, we used 286 WSIs from 68 patients including 1.3 million $480\mu m \times 480\mu m$ ROIs, which is $3588cm^2$ of prostate tissue in total. To the best of our knowledge, this represents the largest validation data set presented thus far in the literature for this problem. All of the validated methods achieved similar or better performance with respect to comparable previously published approaches. The necessary processing times suggest the potential of an optimized, parallel implementation of the algorithms to yield processing speeds compatible with the clinical pathology workflow, upon further multi-centre validation.

2.10.1 Tissue component mapping

The proposed nuclei segmentation method demonstrated robustness to staining variability. In our single-centre data set where manual staining were used, we observed substantial staining variability across and within patients. This was evidenced by the substantial WSI-to-WSI variability in the optimal thresholds on the hematoxylin channel

for accurate nuclei segmentation. This suggests that the concentration of hematoxylin stain for nuclei in WSIs from different patients varies substantially. This is illustrated qualitatively with two example WSIs that were stained very differently (Figure 2.4) and the corresponding hematoxylin channel images (Figure 2.4 (a) middle image sets for both cases). The hematoxylin channel intensity of the nuclei in the more lightly stained image (Figure 2.4 (a) right case) is similar to that of the stroma in the more darkly stained image (Figure 2.4 (a) left case). The proposed method can achieve accurate nuclei segmentation despite the staining variation, as can be observed by comparing the loci and shapes of the nuclei in the output label map images with the original stained images shown in Figure 2.4 (a) and (b). In addition, our method successfully assigned RBCs as stroma/other tissue (Figure 2.4 (c)).

Comparing to machine learning or normalization based approaches [26, 33, 48], our proposed algorithm has lower computational cost, which is important especially for cancer detection throughout entire slides within a reasonable time frame. The algorithm is also case independent, whereas machine learning/normalization-based methods use other images for calibration/training. This suggests the generalization capability of the proposed algorithm.

2.10.2 Cancer vs. non-cancer classification

2.10.2.1 Overall system evaluation

Our system yielded state-of-the-art overall performance (Table 2.3) comparing to literature-reported performance on this problem. Considering we have a large number (1.2 million) of negative samples, our FPRs (Table 2.3) imply a large number of false positives in detection. However, the FPR values (Table 2.3) are small, which also

indicates that the number of false positives is small relative to the total number of negatives. In addition, our FNRs (Table 2.3) imply high recall. More intuitively, in interpreting the confusion matrices (Table 2.5), it is important to note that the total number of negatives in our data set is far larger than the total number of positives. This leads to what appears to be, on first impression, a large number of false positive classifications. However, this concern is tempered when comparing the number of false positives to the generally much larger number of true negatives in each case (i.e. there is a large amount of negative tissue, and the classifiers are correctly identifying it as such most of the time). This is helpful to the clinical scenario where this system is providing assistance in reviewing slides to the pathologist, since a high-recall system minimizes the chance that a cancerous region will be missed and improves pathologist efficiency by drawing attention to most of the cancer on each slide. This is the first study that used all tissues covering WSIs of whole-mount RP sections including all clinically relevant grade groups. This performance suggests the potential to use our proposed pipeline for cancer detection in a clinical setting after multi-centre validation. Deep learning approaches overall outperformed the conventional machine learning based approaches (AUCs in Table 2.3, and the statistical tests in Table 2.4). Also, Tune-AlexNet-RawIM yielded superior overall performance compared to Tune-AlexNet-TCM using the AUC metric (Table 2.3 and Table 2.4).

All the methods were affected by tissue type and the corresponding sample size of each tissue type (Figure 2.7 and Table 2.2). Lower FNRs and FPRs were often associated with larger sample sizes for the different tissue types, while higher error rates were associated with smaller sample sizes, with one anomaly. Although G5+4 has a relatively

large sample size of 8216, its training sample size was small because most of the G5 cancer was concentrated in very few patients, and no tissue from the same patient was used in both training and testing.

The sensitivity of classifier performance to the sample sizes of the different tissue types varies according to the machine learning method used. Tune-AlexNet-RawIM and TCM-Texture-SVM are the most sensitive to sample size, performing worse than the other methods for G5, G4+5, G5+4, and EPE, all of which have relatively small sample sizes (Figure 2.7 and Table 2.2). In comparison, the performances of Tune-AlexNet-TCM and TCM-Texture-FisherC were the least sensitive to the smaller sample sizes of these tissue types. This is reinforced by the observation that in the training sample size experiment, the FNRs of Tune-AlexNet-RawIM and TCM-Texture-SVM decreased much more with increasing sample sizes, compared to the other methods (Figure 2.8, FNR metric). Tune-AlexNet-TCM and Tune-AlexNet-RawIM showed larger fluctuations in FNR compared to the TCM-Texture based methods, when the training sample size is small (Figure 2.8 FNR before red dashed line). We speculate that this may arise due to the inherent randomness associated with fine-tuning AlexNet. As the number of training patients increases, the amplitude of the fluctuation reduces (all error metrics in Figure 2.8, especially FNR), suggesting that larger training sample sizes may increase the stability in performance of the Tune-AlexNet methods. In addition, we found that the amplitude of the FNR fluctuations was larger than the amplitude of the FPR fluctuations. This may be due to the heterogeneity of prostate cancer tissue, such that we have many different cancerous tissue types with relatively smaller samples sizes, compared to the larger samples of non-cancerous tissue.

In the LOPO CV, we also found that, for PIN, the FPRs were smaller when using more complex models (Tune-AlexNet-TCM and -RawIM) (Table 2.3). We speculate that this could be due to PIN's resemblance to cancerous tissue, requiring more complex models or more image information (i.e. raw images or TCMs rather than 14 calculated texture features) to differentiate it from other cancerous tissue. From a clinical perspective, high grade PIN is considered as a putative precursor lesion, as PIN shares features with cancer tissue [49].

2.10.2.2 Performance comparison by methods

Tune-AlexNet-RawIM yielded the highest AUC in all the CV experiments and in the training sample size experiment, but it is more sensitive to sample size. Although Shin et al. [50] have demonstrated efficient (i.e. smaller sample size) training by fine-tuning pre-trained AlexNet, it only performs better than or equal to Tune-AlexNet-TCM and the TCM-Texture based methods with training sets larger than 12 patients (Figure 2.8 after red dotted line, and all CVs in Table 2.3). This is also reflected by the inferior performance of Tune-AlexNet-RawIM on cancerous tissue types involving G5 (see blue bars in Figure 2.7). This could be due to the fact that our data set consisted of two patients having any G5 cancer, combined with the fact that we performed LOPO CV. To illustrate, note the large portion of G5+4 cancerous regions that were missed by Tune-AlexNet-RawIM (white regions in Figure 2.9 (h)). This is of particular concern because overlooking G5 cancer could result in failure to apply adjuvant therapy after surgery, or could increase the pathologist's necessary editing time to correct the cancer maps. Although Tune-AlexNet-RawIM had an error rate of 0% on G5+3 using a large training

sample size of G5 from tissues of G5, G4+4, G4+5, the sample size being tested was too small (9 ROIs from one patient) to be considered representative.

In comparison, Tune-AlexNet-TCM yielded more stable performance in the training sample size experiment and demonstrated superior performance on tissue types with smaller sample sizes. With fewer than 12 training patients, the Tune-AlexNet-TCM yielded a lower FNR than Tune-AlexNet-RawIM in nearly every case (Figure 2.8 before the red dotted line). In LOPO CV, even the samples of G5 involved tissue types were restricted in two patients, lowest FNRs and FPRs were achieved most of the time (Figure 2.7). More intuitively, for example, in Figure 2.9 (g) we can see most of the cancerous regions of G4+5 and G5+4 were captured by Tune-AlexNet-TCM. This suggests that higher-order tissue features (e.g. TCMs) can enhance the performance of Tune-AlexNet when sample size is small. Also, comparing the results between using Tune-AlexNet-TCM and Tune-AlexNet-RawIM in CVs (Table 2.3), the performance differences are negligible. Considering the huge dimensionality reduction from the raw image (227 pixels × 227 pixels × 3 colour channels × 256 intensities per channel) to TCM (227 pixels × 227 pixels × 3 tissue component labels), the computed TCM can effectively reduce the dimensionality, and make salient the key visual cues for the cancer detection problem.

Comparable results between using TCM-Texture-SVM and Tune-AlexNet-TCM and -RawIM in CVs (Table 2.3) suggest that extracting features from TCMs can effectively reduce the high-dimensional image information to 14 TCM-based texture features, and the resulting feature set is appropriate for our problem. This could support better understanding of the key visual cues for our problem. For TCM-Texture based methods, the SVM was more sensitive to sample size than FisherC and LoglC because

(1) its performance monotonically and substantially improved as the training sample size increased, and exceeded the performance of the FisherC and LoglC classifiers with a training sample size of 33 in the training sample size experiment (Figure 2.8 after the black dotted line), and (2) higher FNRs and FPRs were found for tissue types having smaller sample sizes (G5 involved tissue types, excepting PIN as discussed previously) in LOPO CV (Figure 2.7, and Table 2.2).

2.10.3 Limitations

Our results should be interpreted in the context of several limitations of our study. All of our tissues were processed in the same clinical pathology laboratory, and the cancerous annotations were done by one physician and verified by one of the two pathologists. These aspects of our study limit the variability of the material (i.e. the tissues and resulting images) and observers' contours. Also, validations were conducted at the $480\mu m \times 480\mu m$ ROI level with regions containing more than 50% cancer considered as positive. We used backward feature selection, which is a greedy algorithm that may lead to suboptimal performance of the system. Although we used an adaptive threshold for nuclei segmentation, it was a global threshold for each WSI. Although our informal experiments (not reported here) suggested that locally adaptive thresholding did not improve performance, such an approach could be straightforwardly adapted from our methods if needed. Finally, it must be acknowledged that all CV studies may be subject to positive bias in their results; validation using an external data set is required to support clinical translation of this tool.

2.10.4 Conclusion

In conclusion, our proposed methods can automatically map cancerous regions on digitized WSIs of mid-gland RP tissue sections. We validated and compared our proposed methods using the largest pathologist-annotated data reported thus far, with high-precision annotations used as the gold standard. State-of-the-art classification accuracy and speed were achieved. A deep learning approach based on transfer learning with AlexNet using raw image ROIs performed the best overall, outperforming methods based on handcrafted features. However, the deep learning approach was more sensitive to training sample size and was limited in its ability to detect G5 cancer, which is prognostically important. By contrast, training the deep learning system based on our TCMs resulted in less sensitivity to sample size and better detection of G5 cancer. Our proposed adaptive thresholding technique efficiently computes TCMs showing the loci of nuclei, luminal regions, and regions containing stroma and other tissue, compensating for staining variation without any training requirement. This suggests that the proposed 3-class TCM can reduce noise and image complexity while preserving the key information required to enhance the performance of AlexNet when training sample size is small. Upon successful multi-centre validation, this system could support imaging validation studies using annotated histopathology as the gold standard. It could also facilitate quantitative and graphical pathology reporting after RP, which has the potential to support better prognosis, recurrence risk management, and adjuvant therapy planning.

2.11 References

1. Stephenson, A.J., et al., *Predicting the outcome of salvage radiation therapy for recurrent prostate cancer after radical prostatectomy.* J Clin Oncol, 2007. **25**(15): p. 2035-41.

2. Izawa, J.I., *Salvage radiotherapy after radical prostatectomy.* Can Urol Assoc J., 2009. **3**(3): p. 245-250.

3. Gleason, D.F., G.T. Mellinger, and G. Veterans Administration Cooperative Urological Research, *Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. 1974.* J Urol, 2002. **167**(2 Pt 2): p. 953-8; discussion 959.

4. Egevad, L., J.R. Srigley, and B. Delahunt, *International Society of Urological Pathology Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens.* Advances in Anatomic Pathology, 2011. **18**(4): p. 301-305.

5. van der Kwast, T.H., et al., *International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 2: T2 substaging and prostate cancer volume.* Mod Pathol, 2011. **24**(1): p. 16-25.

6. Evans, A.J., et al., *Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens.* Am J Surg Pathol, 2008. **32**(10): p. 1503-12.

7. Epstein, J.I., et al., *The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma.* Am J Surg Pathol, 2005. **29**(9): p. 1228-42.

8. Sun, M., et al., *Insights of modern pathology reports originating from prostate biopsy and radical prostatectomy specimens.* Eur Urol, 2012. **62**(1): p. 40-1.

9. Bettendorf, O., et al., *Implementation of a map in radical prostatectomy specimen allows visual estimation of tumor volume.* Eur J Surg Oncol, 2007. **33**(3): p. 352-7.

10. Eminaga, O., et al., *CMDX(c)-based single source information system for simplified quality management and clinical research in prostate cancer.* BMC Med Inform Decis Mak, 2012. **12**: p. 141.

11. Gibson, E., et al., *Registration of prostate histology images to ex vivo MR images via strand-shaped fiducials.* J Magn Reson Imaging, 2012. **36**(6): p. 1402-12.

12. Soetemans, D.J., *Computer-assisted characterization of prostate cancer on magnetic resonance imaging.* Electronic Thesis and Dissertation Repository, 2017. **4504**.

13. Croke, J., et al., *Proposal of a post-prostatectomy clinical target volume based on pre-operative MRI: volumetric and dosimetric comparison to the RTOG guidelines.* Radiat Oncol, 2014. **9**: p. 303.

14. Mosquera-Lopez, C., et al., *Computer-Aided Prostate Cancer Diagnosis From Digitized Histopathology: A Review on Texture-Based Systems.* IEEE Rev Biomed Eng, 2015. **8**: p. 98-113.

15. Leo, P., et al., *Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images.* J Med Imaging (Bellingham), 2016. **3**(4): p. 047502.

16. Boyce, B.F., *Whole slide imaging: uses and limitations for surgical pathology and teaching.* Biotech Histochem, 2015. **90**(5): p. 321-30.

17. Magee, D., et al. *Colour normalisation in digital histopathology images*. in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*. 2009. Daniel Elson.

18. Basavanhally, A. and A. Madabhushi, *EM-based segmentation-driven color standardization of digitized histopathology*. SPIE Medical Imaging. Vol. 8676. 2013: SPIE.

19. Mosquera-Lopez, C. and S. Agaian, *Iterative local color normalization using fuzzy image clustering*. SPIE Defense, Security, and Sensing. Vol. 8755. 2013: SPIE.

20. Tabesh, A., et al., *Multifeature prostate cancer diagnosis and Gleason grading of histological images.* IEEE Trans Med Imaging, 2007. **26**(10): p. 1366-78.

21. Farjam, R., et al., *An image analysis approach for automatic malignancy determination of prostate pathological images.* Cytometry B Clin Cytom, 2007. **72**(4): p. 227-40.

22. Tahir, M.A. and A. Bouridane, *Novel round-robin tabu search algorithm for prostate cancer classification and diagnosis using multispectral imagery.* IEEE Trans Inf Technol Biomed, 2006. **10**(4): p. 782-93.

23. Bouatmane, S., et al., *Round-Robin sequential forward selection algorithm for prostate cancer classification and diagnosis using multispectral imagery.* Machine Vision and Applications, 2011. **22**(5): p. 865-878.

24. Sun, X., et al., *Automatic diagnosis for prostate cancer using run-length matrix method*. SPIE Medical Imaging. Vol. 7260. 2009: SPIE.

25. Yu, E., et al. *Detection of prostate cancer on histopathology using color fractals and Probabilistic Pairwise Markov models*. in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2011.

26. Gorelick, L., et al., *Prostate histopathology: learning tissue component histograms for cancer detection and classification.* IEEE Trans Med Imaging, 2013. **32**(10): p. 1804-18.

27. Peyret, R., et al., *Automatic classification of colorectal and prostatic histologic tumor images using multiscale multispectral local binary pattern texture features and stacked generalization.* Neurocomputing, 2018. **275**: p. 83-93.

28. Kwak, J.T. and S.M. Hewitt, *Nuclear architecture analysis of prostate cancer via convolutional neural networks.* IEEE Access, 2017. **5**: p. 18526-18533.

29. Doyle, S., et al., *A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies.* IEEE Trans Biomed Eng, 2012. **59**(5): p. 1205-18.

30. Litjens, G., et al., *Automated detection of prostate cancer in digitized whole-slide images of H and E-stained biopsy specimens*. SPIE Medical Imaging. Vol. 9420. 2015: SPIE.

31. Litjens, G., et al., *Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis.* Sci Rep, 2016. **6**: p. 26286.

32. Monaco, J.P., et al., *High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models.* Med Image Anal, 2010. **14**(4): p. 617-29.

33. Rashid, S., et al., *Automatic pathology of prostate cancer in whole mount slides incorporating individual gland classification.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2019. **7**(3): p. 336-347.

34. Epstein, J.I., et al., *The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System.* Am J Surg Pathol, 2016. **40**(2): p. 244-52.

35. DiFranco, M.D., et al., *Ensemble based system for whole-slide prostate cancer probability mapping using color texture features.* Comput Med Imaging Graph, 2011. **35**(7-8): p. 629-45.

36. Nguyen, K., A.K. Jain, and B. Sabata, *Prostate cancer detection: Fusion of cytological and textural features.* J Pathol Inform, 2011. **2**: p. S3.

37. Nir, G., et al., *Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts.* Med Image Anal, 2018. **50**: p. 167-180.

38. Han, W., et al., *Automatic cancer detection and localization on prostatectomy histopathology images*. SPIE Medical Imaging. Vol. 10581. 2018: SPIE.

39. Haralick, R.M., K. Shanmugam, and I. Dinstein, *Textural Features for Image Classification.* IEEE Transactions on Systems, Man, and Cybernetics, 1973. **SMC-3**(6): p. 610-621.

40. Galloway, M.M., *Texture analysis using grey level run lengths.* Computerized Medical Imaging and Graphics, 1975. **4**: p. 172-179.

41. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

42. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.

43. Ruifrok, A.C. and D.A. Johnston, *Quantification of histochemical staining by color deconvolution.* Anal Quant Cytol Histol, 2001. **23**(4): p. 291-9.

44. Jahne, B. and B. Jaehne, *Practical handbook on image processing for scientific applications*. 1995: CRC Press, Inc.

45. Runge, C., *Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten [On empirical functions and the interpolation between equidistant ordinates].* Zeitschrift Mathematische Physik, 1901. **46**: p. 224-43.

46. Horn, R.A. and C.R. Johnson, *Matrix Analysis*. 2 ed. 2012, Cambridge: Cambridge University Press.

47. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

48. Khan, A.M., et al., *A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution.* IEEE Trans Biomed Eng, 2014. **61**(6): p. 1729-38.

49. Epstein, J.I., *The lower urinary tract and male genital system.* Robbins and Cotran Pathologic Basis of Diseases, 2005.

50. Shin, H.C., et al., *Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning.* IEEE Trans Med Imaging, 2016. **35**(5): p. 1285-98.

Chapter 3

A version of this chapter has been submitted to Scientific Reports for publication and is currently under review: Wenchao Han, Carol Johnson, Mena Gaed, Jose A. Gomez-Lemus, Madeleine Moussa, Joseph Chin, Stephen Pautler, Glenn Bauman, and Aaron Ward, "Histologic tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens."

3 Histologic tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens

3.1 Introduction

The most used treatment for prostate-cancer (PCa) that is organ-confined is radical prostatectomy (RP), the removal of the prostate gland. Approximately 40% of prostate cancer patients undergo this surgery each year in the United States [1]. Serum prostate-specific antigen (PSA) relapse occurs in 17%–29% of patients, reflecting cancer recurrence [2, 3]. Post-surgery prognosis, recurrence prediction, and selection and guidance for adjuvant therapy are all informed by the surgical pathology report. Typical pathology reports include tumour size, location, spread, and aggressiveness levels. In addition, PCa patients are grouped based on the Gleason score (GS), which is computed as the sum of the primary and secondary Gleason grades [3] at RP, into grade group 1 (GS 6; G3+3), grade group 2 (GS 7; G3+4), grade group 3 (GS7; G4+3), grade group 4 (GS 8; G4+4) and grade group 5 (GS 9–10; G4+5, G5+4, and G5+5) disease [4, 5], with treatment determined according to the risk level [6]. Thus, although accurate post-RP risk stratification is crucial, currently, clinical pathology reporting is primarily qualitative and subject to intra- and inter-observer variability. This leads to challenges for quantitative

and repeatable pathology reporting and interpretation regarding the lesion size, location, spread, and Gleason grade or score [3, 7-10].

Whole-mount tissue sections, where the entire cross section of tissue from the gross section is mounted to the slide, give the pathologist a better overview to facilitate the identification of multiple tumour foci [11]. If cancerous regions of interest (ROI) could be accurately and precisely contoured on whole mount WSIs of RP sections, this would enable quantitative reporting of tumour size, location, and grade. This would yield quantitative clinical pathology reporting and would benefit research studies, including imaging validation studies, which require an annotated histologic gold standard [12-14]. However, such manual contouring is too time consuming to perform as part of a routine clinical workflow, and is resource-intensive when performed as part of research studies. There is therefore an unmet need for an approach that can detect and grade cancerous regions accurately and quickly on digitized whole-mount histopathology images of RP tissue sections.

Many published methods have demonstrated the potential of machine learning approaches for automatic prostate cancer detection and grading on digital histopathology images [15]. High-resolution digital histopathology images acquired from RP specimens contain a large number of pixels; for instance, a typical whole-mount image of the mid-gland can contain more than four gigapixels. Consequently, most published work performs validation using a small subset of selected regions of interest (ROIs) to reduce computational demands [15]. A few studies [16-22] have worked on cancer detection using whole-slide-images (WSIs). Doyle et al. [16] and Litjens et al. [17, 18] have demonstrated the ability to process WSIs of much smaller biopsy tissues for finding

prostate cancer using automatic systems. Monaco et al. [19] and Rashid et al. [22] have demonstrated cancer detection systems for finding prostate cancer on WSIs of RP tissue sections with practical processing times by classifying segmented glands, but they reported limitations regarding detection of high-grade cancer tissue using their methods. DiFranco et al. [20] and Nguyen et al. [21] have tested their methods on the WSIs of RP tissue sections, but the sample sizes were 14 patients and 11 WSIs for the two studies, respectively. For grading, Nir et al. [23] validated on the largest number of tissue samples from the tissue micro arrays (TMAs) of RP tissue sections.

Comprehensive validation using all available tissue covering all clinically relevant grade groups avoids bias due to ROI selection and tests the system against the full variability in terms of staining and cancerous tissue appearance. It is also important to ensure that in cross-validation, samples are chosen such that the training and testing sets do not contain samples from the same patient. This is particularly important considering 1) the heterogeneous patterns of each grade [3], 2) the similar patterns among different grades, 3) the large staining variability among WSIs [24, 25], and 4) the requirement for practical processing times for clinical translation to the pathology laboratory.

In recent years, deep learning has demonstrated potential for analyzing digital histology images. For example, Litjens et al. [18] used deep learning to find prostate cancer on biopsy tissues. Kwak et al. [26] used deep learning to classify ROIs from TMAs of RP tissue sections as cancerous vs. non-cancerous. However, the use of deep learning in finding and, in particular, grading prostate cancer is still new. In addition, in the previous studies, semantic features (i.e. higher-level tissue components such as

nuclei, lumen, etc.) have been demonstrated as crucial factors for finding and grading prostate cancer as they reflect the differentiation of cancerous tissue [27]. Many studies used features extracted from semantic feature maps and reported promising results [15]. However, the importance and applicability of those methods were not fully evaluated due to lack of comprehensive comparisons of system performance for detecting and grading PCa, especially validating on mid-gland whole-mount WSIs of RP sections.

In this study, we investigated the utility of tissue components (specifically, nuclei, lumen, and stroma/other tissue) as cues used in 7 different machine learning approaches (3 non-deep learning and 4 deep learning) for finding and grading prostate cancer on whole-mount WSIs of RP sections, validating on all available ROIs covering each WSI in a data set of 286 WSIs from 68 RP patients.

Of the 299 whole-mount WSIs of mid-gland tissue sections obtained from 71 RP surgical specimens using a standard protocol at our local centre [27], 13 WSIs from 3 patients were used for system tuning and the remaining 286 WSIs from 68 patients were used for validation. After digitization, each WSI was annotated at 20X by our trained physician with each tumour contoured and the grade indicated by contour colour (Figure 3.1 and Figure 3.2). The annotations were verified by one of two genitourinary pathologists. Each WSI was partitioned into a set of ROIs with sizes of 480µm × 480µm.

Figure 3.1 describes the training of the system. We assigned each ROI a tissue type label (i.e. cancer or non-cancer, and the Gleason grade for cancerous ROIs) based on the manual annotations done by the expert. We labeled each image pixel as one of three classes: nuclei, lumen, and stroma/other using our previously proposed method [28], to

generate three-class tissue component maps (TCMs). We also used the same technique to generate simpler binary maps: nuclei maps and lumen maps. We trained the system with 7 different machine learning approaches, enumerated as follows: 3 conventional machine learning approaches: (1) a Fisher linear discriminant classifier (FisherC), (2) a logistic linear classifier (LoglC), (3) a support vector machine classifier (SVM) with calculated texture features extracted from the TCMs, and 4 deep learning approaches via fine-tuning of AlexNet[29] with the (4) nuclei maps (AlexNet-Nuclei), (5) lumen maps (AlexNet-Lumen), (6) three-class TCMs (AlexNet-TCM), and (7) raw image ROIs (AlexNet-RawIM).

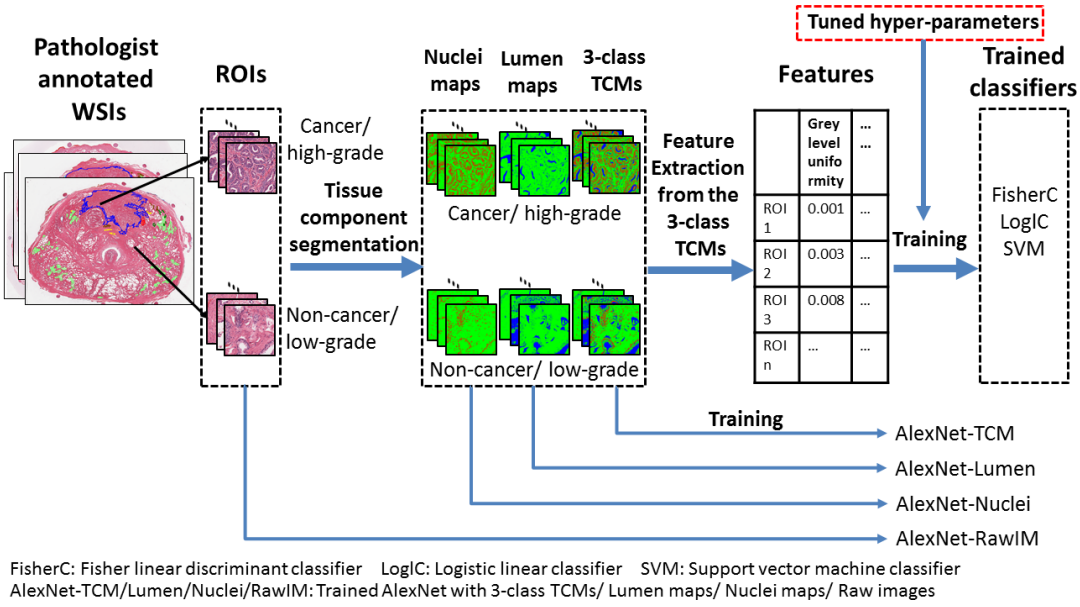
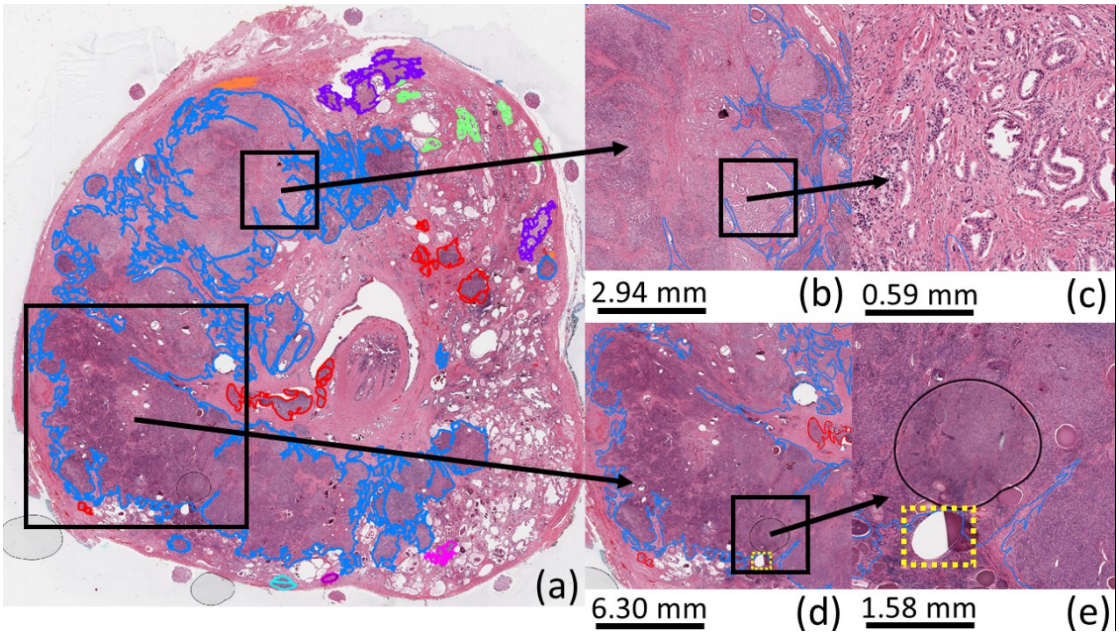


Figure 3.1: Pipeline for system training for cancer vs. non-cancer classification or high- vs. low-grade classification. For tissue component maps, nuclei are labeled in red, luminal regions are labeled in blue, and stroma or other tissue components are labeled in green.



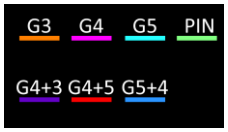


Figure 3.2: WSI of H&E stained histology prostate tissue. (b), (c), (d), and (e) are zoomed from the black square highlighted regions from the WSI. (d) and (e) show a region of torn tissue (yellow dashed square) and a region of poor focus (circle).

We performed 3 experiments for our cancer detection and grading problems: classifying all relevant ROIs as 1) cancer vs. non-cancer, 2) high-(G4) vs. low-(G3) grade cancer, 3) high-(G4 & G5) (i.e., G4, G5, G4+5, G5+4) vs. low-(G3) grade cancer. For experiment 2), ROIs containing ≥50% G4 cancer were considered as high-grade, and ≥ 50% G3 as low-grade. For experiment 3), ROIs containing ≥50% G4 and G5-involved (i.e., G4, G4+5, G5, G5+4), denoted as G4 & G5, cancer were considered as high-grade, and ≥ 50% G3 as low-grade. Since G4+3 and G3+4 cancer have both high- and low-grade cancer tissue, we used those tissue samples for cancer detection but not for grading experiments. The validations were conducted using all available ROIs for each WSIs using leave-one-patient-out (LOPO) cross-validation (CV), during which training and

testing ROIs were never drawn from the same patient. We measured cumulative error metrics of error rate, false negative rate (FNR), false positive rate (FPR), and area under the receiver operating characteristic curve (AUC), comparing the predicted label from each machine learning technique for each ROI with the reference standard label assigned to the ROI based on the pathologist's annotations. We also measured the error rate for each tissue type separately using each of our seven approaches. Our implementation used Matlab 2018a (The Mathworks, Natick, MA), OpenCV 3.1 for SVM implementation, and PRtools 5.0 (Delft Pattern Recognition Research, Delft, The Netherlands) for implementation of FisherC and LoglC machine learning algorithms.

Table 3.1 Cumulative error metrics for cancer vs. non-cancer and high-vs. low-grade cancer classifications from leave-one-patient-out cross-validation. G4 vs. G3: high-(G4) vs. low-(G3) grade classification. G4 & G5 vs. G3: high-(G4 & G5) vs. low-(G3) grade classification. Bolded number: highest AUC in the experiment across 7 different methods.

	FisherC	LoglC	SVM	AlexNet-RawIM	AlexNet-TCM	AlexNet-Nuclei	AlexNet-Lumen
Cancer vs. non-cancer							
Error rate	13.50%	12.20%	8.60%	5.90%	6.10%	9.00%	13.20%
FNR	14.90%	16.90%	19.90%	19.20%	15.10%	15.80%	21.00%
FPR	13.50%	12.00%	8.20%	5.50%	5.80%	8.80%	13.00%
AUC	0.927	0.926	0.928	0.957	**0.964**	0.937	0.896
G4 vs. G3							
Error rate	20.40%	20.00%	21.90%	11.40%	12.70%	13.90%	25.90%
FNR	26.20%	27.00%	38.00%	24.00%	28.90%	32.30%	61.80%
FPR	18.90%	18.20%	17.80%	8.20%	8.60%	9.20%	16.70%
AUC	0.858	0.85	0.783	**0.934**	0.904	0.891	0.654
G4 & G5 vs. G3							
Error rate	20.40%	20.90%	26.20%	16.90%	13.20%	15.20%	35.30%
FNR	33.10%	32.60%	43.90%	27.40%	20.00%	19.80%	52.90%
FPR	9.30%	10.80%	10.80%	7.60%	7.30%	11.20%	20.10%
AUC	0.886	0.875	0.815	0.916	**0.923**	0.919	0.66

3.2 Results

3.2.1 Prostate cancer detection

The quantitative results for cancer vs. non-cancer classification from our LOPO CV using each method are reported in Table 3.1. All methods yielded AUCs higher than

0.92 except AlexNet-Lumen, which has an AUC of 0.896. AlexNet-TCM yielded the highest AUC of 0.964 (bolded in Table 3.1). AlexNet-RawIM and AlexNet-Nuclei yielded the second- and third-highest AUCs of 0.957 and 0.93 respectively. In general, the methods of fine-tuning AlexNet have higher AUCs and much lower FPR than the conventional machine learning methods.

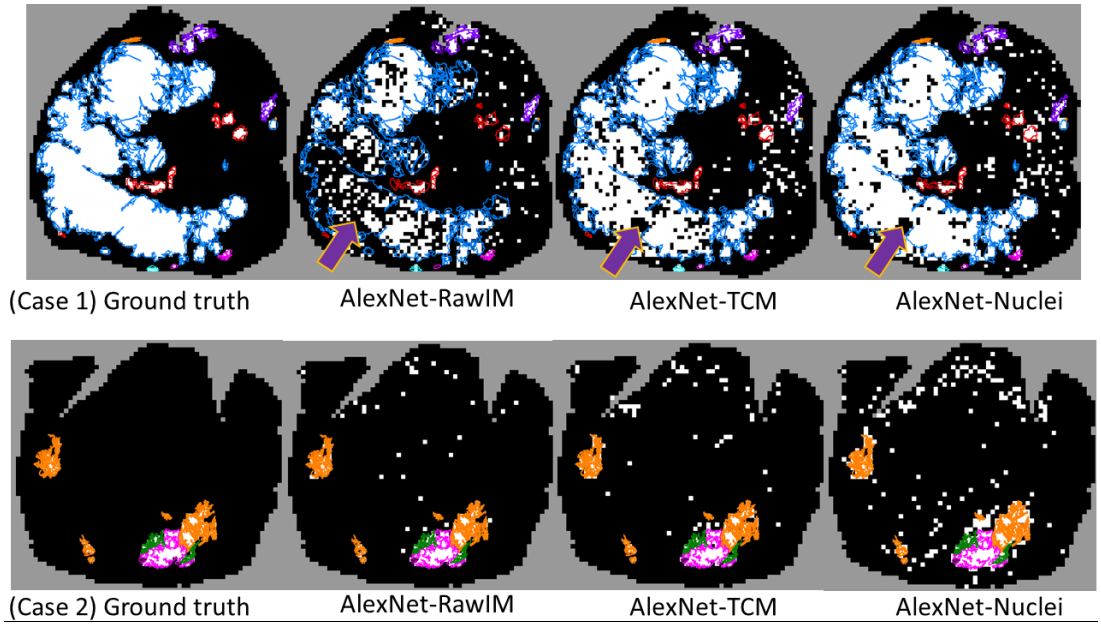
Table 3.2: Number of ROIs for each tissue type.

	Tissue types	Sample size
Cancerous ROIs	G3	14719
	G3+4	6008
	G4+3	3839
	G4	3949
	G4+5	725
	G5+4	8216
	G5	37
	G5+3	16
	EPE	272
Non-cancerous ROIs	Atrophy	5433
	PIN	26449
	Healthy tissue/BPH	1178814

Figure 3.3 shows our system's mapping of cancer throughout entire WSIs for two samples cases. The major cancerous and non-cancerous regions were correctly labeled by the systems for both cases. In case 1, AlexNet-TCM and AlexNet-Nuclei have similar results, while AlexNet-RawIM performs the worst, with many more false negatives in the G5+4 cancerous region. Figure 3.2 shows the original H&E stained WSI of case 1. The

bottom two images are the zoomed in view from the square highlighted region from the WSI. It includes an unfocused region, and a region with torn tissue (yellow dashed square highlighted region in Figure 3.2 (e)). For those regions, all methods falsely classified them as negatives (regions indicated by purples arrows in Figure 3.3 (case 1)). In case 2, AlexNet-RawIM and AlexNet-TCM have similar results, while AlexNet-Nuclei has more false positives. The major cancerous regions are G3, G3+4, and G4.

We calculated the FNRs for cancerous ROIs and the FPRs for non-cancerous ROIs, effectively computing the error rates for each of these tissue types. The results of the LOPO CV experiments are shown in Figure 3.4. Table 3.2 shows the number of ROIs used for each tissue type. G5, G4+5, G5+4, and EPE yielded higher error rates. Table 3.2 and Figure 3.4 demonstrate that in general, higher error rates corresponded to smaller sample sizes; G5+4 was the exception. For those tissue types, with the exception of EPE, AlexNet-Nuclei, AlexNet-TCM, and FisherC yielded much lower error rates than the other methods. Among those methods, AlexNet-Nuclei has the lowest error rate. For G4-involved tissue types (i.e. G3+4, G4+3, and G4), FisherC yields the lowest error rates, and LoglC achieved similar performance. For other tissue types, AlexNet-RawIM yielded the lowest error rates. Those tissue types are primarily non-cancerous and G3 cancerous tissues, and they have larger sample sizes.



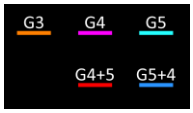


Figure 3.3: Cancer maps generated by each of the trained systems. White: cancerous tissue regions. Black: non-cancerous tissue regions. Colour contours: pathologist manual annotations. The purple arrows point to unfocused areas and areas with torn tissue as indicated in Figure 3.2 (d, e).

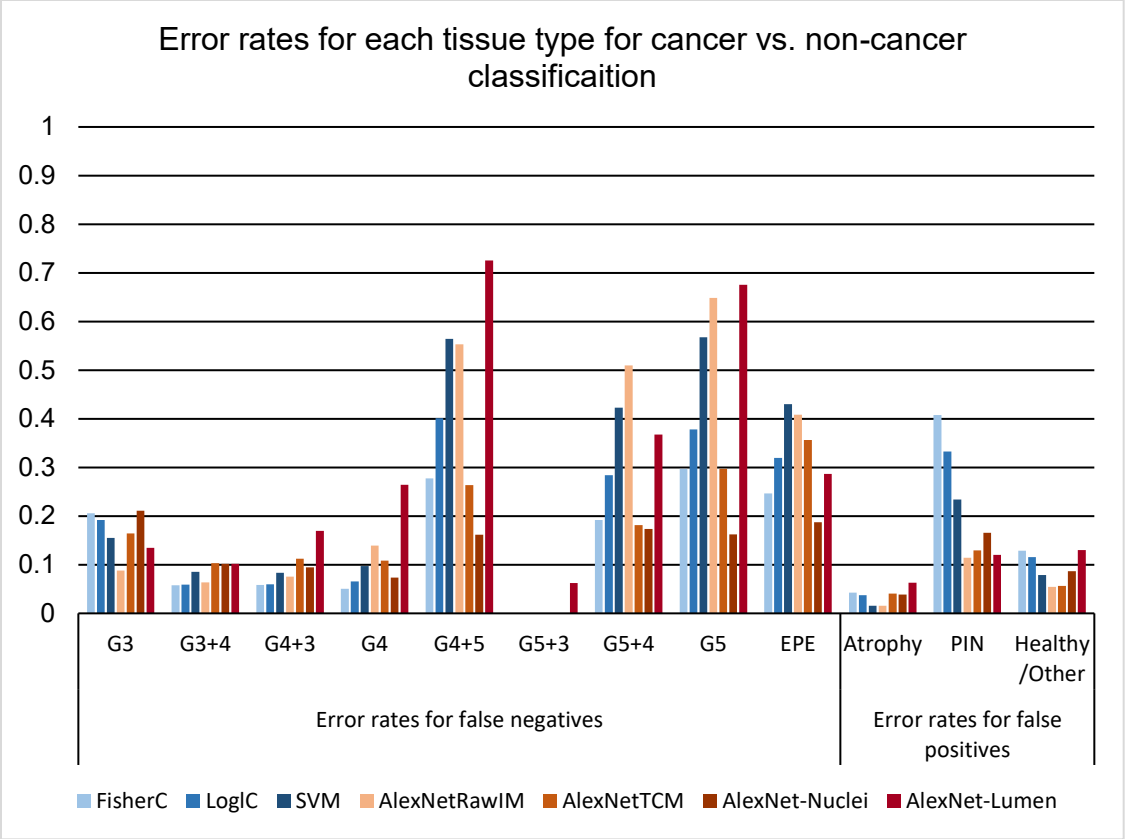


Figure 3.4: FNR for cancer tissue types, and FPR for non-cancer tissue types to reflect the error rate for each tissue type, for each classifier from leave-one-patient-out cross-validation of cancer vs. non-cancer classification.

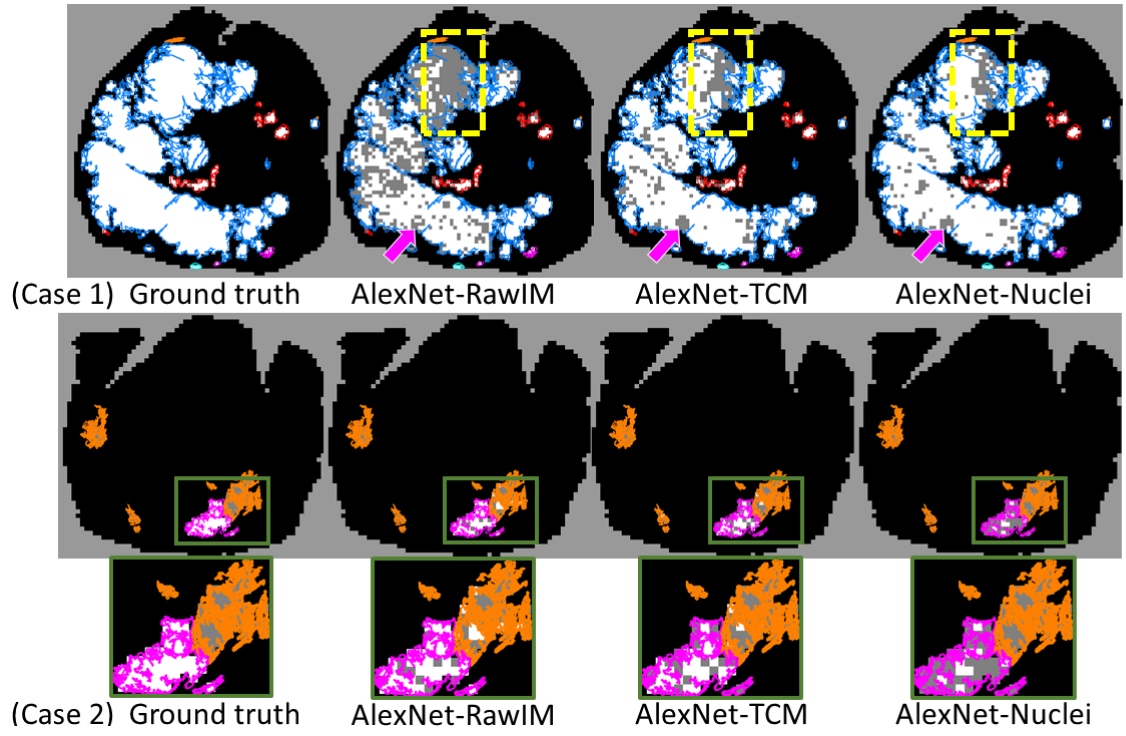
3.2.2 Prostate cancer grading (high- vs. low- grade)

The quantitative results for high- vs. low-grade cancer classification from our LOPO CV using each method are reported in Table 3.1. For high-(G4) vs. low-(G3) grade classification, AlexNet-RawIM yielded the highest AUC of 0.934, followed by AlexNet-TCM and AlexNet-Nuclei with AUCs of 0.904, and 0.891 respectively. For high-(G4 & G5) vs. low-(G3) grade classification, AlexNet-TCM, AlexNet-Nuclei and AlexNet-RawIM are the top three performing methods with AUCs of 0.923, 0.919, and 0.916 respectively. SVM and AlexNet-Lumen had much lower AUCs than other methods

for both of the experiments. Except for AlexNet-Lumen, methods of fine-tuning AlexNet yielded higher AUCs, lower FPRs and FNRs than the conventional machine learning approaches for both experiments.

In Figure 3.5, two samples of whole-slide mapping of graded cancer are shown. The major cancerous regions are correctly graded and labeled by the systems. For case 1, similar to cancer detection, AlexNet-TCM and AlexNet-Nuclei yield similar performance, and AlexNet-RawIM has many more false negatives. For the unfocused and torn tissue regions (Figure 3.2 (d) and (e)) and the regions with lower Gleason patterns (Figure 3.2 (b) and (c)), all methods incorrectly labeled the tissue as negative (regions highlighted with a yellow dashed square in Figure 3.5). In case 2, AlexNet-TCM and AlexNet–RawIM yielded similar performance, while AlexNet-Nuclei had more false negatives (regions highlighted with a green square in the zoomed in view in Figure 3.5).

From the LOPO CV experiments for high-(G4 &G5) vs. low-(G3) classification, we calculated the error rates for each tissue type (i.e. taking high-grade cancer as "positive" in these experiments, we calculated FNRs for high-grade cancer tissues types, and the FPRs for the low-grade cancer tissue types). The results are shown in Figure 3.6. We found higher error rates for each of the high-grade cancer tissue types, compared to the error rate for the low-grade cancer tissue type. For tissue types which have G5 cancer tissue involved, AlexNet-Nuclei yielded the lowest error rates. For all other tissue types, AlexNet-TCM had the lowest error rates.



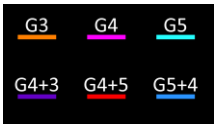


Figure 3.5: Label maps for high- vs. low-grade cancer grading generated by each of the trained systems. White: high-grade cancerous tissue regions. Grey: low-grade cancerous tissue regions. Black: tissue section. Colour contours: pathologist's manual annotations. The region highlighted by the yellow square refers to the tissue regions in Figure 3.2 (b, c). The region indicated by the pink arrow refers to the unfocused areas and regions with torn tissue in Figure 3.2 (d, e)

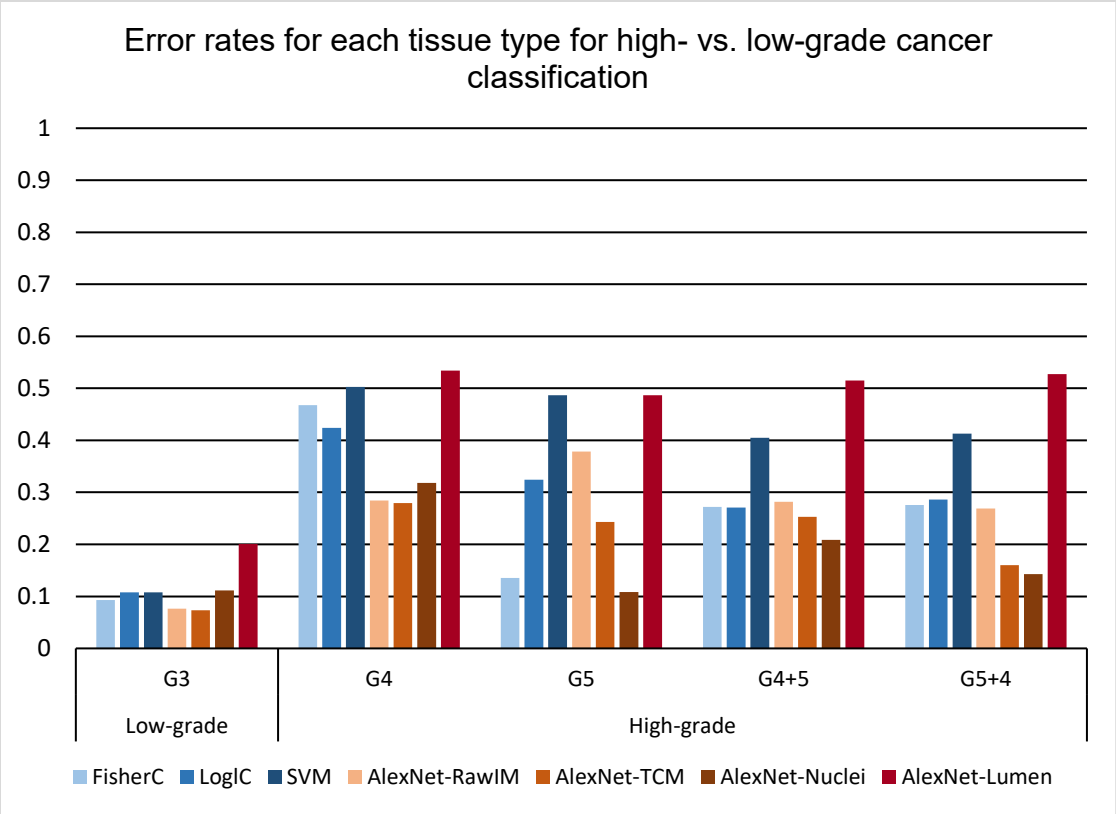


Figure 3.6: Error rate (FNR for high grade cancer, FPR for low-grade cancer)) for each tissue type for each classifier from leave-one-patient-out cross-validation of high-(G4 & G5) vs. low-(G3) grade classification.

3.3 Discussion

Although using machine learning to analyze H&E histology images for prostate cancer detection and grading is an active research field, there are relatively few studies validating on whole-mount RP tissue sections, and the use of deep learning for this problem is still relatively new [18]. In addition, many studies have demonstrated that tissue component features are important for prostate cancer detection and grading [15], but the effects of those tissue components on system performance for cancer detection and grading for different types of tissue were not directly compared. Therefore, in our

study we used different machine learning approaches with different tissue component maps, and compared the performances for both the cancer detection and grading problems on the largest expert-annotated dataset of RP tissue sections reported thus far.

In general, for both cancer detection and grading, AlexNet-TCM achieved the best overall performance, followed closely by AlexNet-RawIM. Conventional machine learning approaches demonstrated inferior but comparable overall performance (AUCs in Table 3.1). This suggests that the 3-class TCMs provide a set of major cues for prostate cancer detection and grading. This is also reflected by very similar performance of AlexNet-RawIM to that of AlexNet-TCM. The observed slightly inferior overall performance by using raw images could be due to irrelevant or redundant information (e.g. red blood cells) from the raw images, resulting in confounders to which the network could overfit.

With the exception of AlexNet-Lumen, fine-tuning AlexNet based approaches achieved better performance than the conventional machine learning based approaches for both cancer detection and grading. This suggests fine-tuning AlexNet outperforms conventional machine learning based approaches overall. This can also be supported by direct comparison of AlexNet-TCM and conventional machine learning approaches. For cancer detection, AlexNet-TCM had lower error rates for most tissue types (Figure 3.4), therefore better overall performance was achieved. For grading, it had higher AUCs for the two grading experiments (Table 3.1). For the second experiment, with the exception of G5 involved cancer, AlexNet-TCM had lower error rates for all tissue types (Figure 3.6). The worst performance was yielded using AlexNet-Lumen, suggesting that the lumen maps provide insufficient information for our problem. This is also suggested by

the much larger performance differences between AlexNet-Lumen and other methods for cancer grading, compared to the performance differences for cancer detection (Table 3.1). We speculate that because the tissue appearances are much more similar for cancerous tissues of different grades than for cancer vs. non-cancerous tissues, more tissue information is needed for cancer grading.

The performances of all the machine learning methods we used are sensitive to sample size, with sensitivity varying according to machine learning method used, classification task, and tissue type. Lower error rates are usually associated with larger sample sizes, and vice versa, except for G5+4 cancer tissue (Table 3.2, Figure 3.4, and Figure 3.6). G5+4 appears to be an exception with a sample size of 8,216, which is relatively large. However, most of the G5 involved cancer (including G5+4) occurred in a small number of patients. Since we used LOPO CV, tissue from a single patient never appeared in both the training and testing sets, reducing the number of occurrences of G5 cancer in training. For cancer detection, AlexNet-RawIM was the most sensitive to the sample size, while AlexNet-TCM and AlexNet-Nuclei, and the conventional machine learning approaches (except for SVM) were less sensitive to sample size (Figure 3.4 and Table 3.2). This suggests higher-order semantic features (e.g. tissue component based features) can improve the robustness of the system to smaller training sample sizes. However, for cancer grading, this was not the case for G4 cancer tissues (Figure 3.6), where conventional machine learning based approaches had substantially higher error rates than the AlexNet-based approaches, compared to other tissue types. We speculate that this could be due to the relative similarity of the G3 and G4 patterns, requiring more complex deep learning model to differentiate them.

For both cancer detection and grading, AlexNet-Nuclei achieved similar but slightly inferior overall performance to AlexNet-TCM and AlexNet-RawIM (Table 3.1), and the best performance for high-grade cancer tissue types (Figure 3.5 and Figure 3.6). This suggests that among the 3-class TCMs, the nuclei maps capture the key cues for our problems, especially for higher-grade cancer tissue types (i.e. G5, G4+5, G5+4). Adding lumen features (using AlexNet-TCM) or other features (using AlexNet-RawIM) improved the performance for most tissue types, but not for higher-grade cancer tissue types (Figure 3.5 and Figure 3.6). This also makes sense from the clinical pathology perspective. Since higher grade cancer tissue (G5-involved cancer tissues) are poorly differentiated with merged glands and much less stroma tissue (Figure 3.2), luminal and stroma features are not helpful for identifying those tissue types. Also, those tissues have larger amounts of nuclei, which leads to darker hematoxylin stain (Figure 3.2). Thus, the 3-class TCMs and raw images are likely to contain more extraneous information, compared to the nuclei maps. Vice-versa, this explains better performance for the G3-involved and non-cancerous tissue types (Figure 3.5 and Figure 3.6, and Table 3.1, G4 vs. G3) using raw images, and consistent performance across all tissue types using TCMs.

Kwak et al. [26] have also previously reported that a nucleus seed map is essential for prostate cancer detection using machine learning techniques. On a data set consisting of 707 sample cores from 4 TMAs, they found that nucleus seed maps trained with their proposed convolutional neural network (CNN) yielded better performance than raw images trained with other CNNs (including AlexNet). The use of different data sets and sample sizes in their study may explain the differences with respect to our results.

The results of this study must be interpreted in the context of its limitations. First, all of the tissue sections were processed in one clinical pathology facility. Since tissue processing conditions and protocols vary from centre to centre, multi-centre studies are needed to translate these techniques to practice. We would expect these issues to affect the methods using raw images more than those using TCMs, the computation of which is adaptive and calibration-free. Second, since our study was validated using annotations made by one physician and verified by one of two genitourinary pathologists, measurement of the impact of inter-pathologist assessment variability not within the scope of this study. Third, our conventional machine learning methods may yield sub-optimal performance due to the following reasons: 1) we only investigated first- and second-order statistical features for texture feature quantization, 2) backward feature selection is a greedy algorithm, and 3) there exist many types of classifiers that were not tested in our study.

In conclusion, this work demonstrated automatic prostate cancer detection and grading on gigapixel WSIs of RP tissue sections using machine learning approaches with the state-of-the-art performance and practical processing time, and testing on the largest amount of expert annotated tissue so far. Fine-tuning pre-trained AlexNet demonstrated better performance than conventional machine learning based approaches overall. We found that the 3-class TCMs captured the main information for both prostate cancer detection and grading, and yielded robust performance across different tissue types and sample sizes. The best overall performance was achieved using the 3-class TCMs with transfer learning using AlexNet. In the 3-class TCMs, the nuclei maps provided the most important information overall, and was essential for classifying G5-involved cancerous

tissue types for both cancer detection and grading. Future work could include detection and quantification of tissue margin involvement and other prognostic pathology features.

3.4 Methods

3.4.1 Data

3.4.1.1 Materials and imaging

We obtained informed consent from all 71 patients in our study, and this study was approved by our institutional Human Subjects Research Ethics Board. All experiments were performed in accordance with relevant guidelines and regulations. All patients had biopsy-confirmed prostate cancer, clinical stage T1 or T2. From these patients we obtained 299 H&E-stained, 4μm thick, paraffin-embedded mid-gland tissue sections, and acquired a whole-slide image from each. We used the same protocol as described in our previous paper [27] and processed all tissues in our clinical pathology laboratory. We used two different scanners to obtain images at 20X (0.5μm/pixel) in bigtiff pyramid format without compression: an Aperio ScanScope GL (Leica Biosystems, Wetzlar, Germany) for sections from 46 patients and an Aperio ScanScopeAT Turbo (Leica Biosystems, Wetzlar, Germany) for sections from the other 25 patients. This process yielded 24-bit RGB colour images at 0.5μm/pixel.

3.4.1.2 Manual annotation

A trained physician (Gaed) contoured and graded each WSI at 20× magnification using a Cintiq 12WX pen-enabled display (Wacom Co. Ltd., Saitama, Japan) with the ScanScope ImageScope v11.0.2.725 image viewing software (Aperio Technologies, Vista, CA, USA) [3]. Each contour was verified, and edited as necessary, by one of two

genitourinary pathologists (Moussa or Gomez). The zoomed region in Figure 3.2 demonstrates the level of precision of our contouring.

3.4.1.3 Ground truth ROI labeling

We separated each WSI into a set of square 960×960 pixel ROIs. We assigned each ROI a label according to the manual pathology annotations with 50% threshold. For cancer detection, ROIs containing more than 50% cancerous tissue were considered cancerous; all other ROIs were considered non-cancerous. Non-cancerous regions contained confounders such as atrophy, benign prostatic hyperplasia (BPH), high-grade prostate intraepithelial neoplasia (PIN), and inflammation. For cancer grading, ROIs containing more than 50% high-grade cancer tissue were considered positive, otherwise negative. The sample size of each tissue type is shown in Table 3.2.

3.4.2 Data separation for system tuning and feature selection

We used a "tuning data set" of 13 WSIs from 3 patients for hyper-parameter tuning and feature selection. The tuning data set was entirely separate from the rest of the data and was not used for cross validation. We used the 68 remaining patients for cross-validation. WSIs from both scanners were included in both the tuning and cross-validation data sets.

3.4.3 Tissue component mapping

Tissue staining makes salient tissue components having semantic meaning to the pathologist. We used our previously developed methods [28] to assign a label to each image pixel to generate (1) a nuclei map, (2) a lumen map, (3) a 3-class TCM for further analysis, via (1) segmentation of nuclei using colour deconvolution [30] and our

previously proposed adaptive thresholding algorithm [28] to generate the nuclei map, (2) segmenting luminal areas by global thresholding in the red-green-blue (RGB) colour space to generate the lumen map, and (3) combining the results of nuclei and lumen segmentation and designating the rest pixels as “other” to generate the 3-class TCM. The details of these methods are described as follows.

3.4.3.1 Nucleus mapping

We separated the H&E stains into three image channels of hematoxylin stain, eosin stain, and the background, using a colour deconvolution algorithm [30]. We applied this algorithm to each ROI independently using the standard deconvolution matrix used by Ruifrok and Johnston [30], which separated each ROI into three grey-level images corresponding to the amount of hematoxylin, eosin, and background respectively. Most substances within nuclei bind to hematoxylin since they are basophilic. Therefore we used the hematoxylin channel for nuclei segmentation by adaptive thresholding [28].

There are large staining differences across different images from different patients even if the tissue sectioning, staining and scanning were performed using consistent protocols in the same laboratory. The grey-level intensity variation in the hematoxylin channel across different WSIs, which results from staining variability, makes global thresholding not applicable for nucleus segmentation. We therefore used our previously proposed adaptive thresholding method [28]. For each WSI, the segmentation threshold was selected based on a cumulative assessment of 2,000 randomly-selected $120\mu m \times 120\mu m$ ROIs to lie within the prostate (i.e. to avoid clear slide areas) and to not contain tissue marking dye (i.e. to avoid areas of artefact). This makes the segmentation threshold specific to each WSI of the RP tissue section, therefore compensating for the staining

variability across different WSIs. The thresholds for sample selection were derived by inspection of the tuning data set only.

This process will sometimes mislabel RBCs as nuclei because, like nuclei, RBCs stain with hematoxylin. We can distinguish RBCs from nuclei based on the fact that RBCs have higher red-pink saturation. We computed a cumulative histogram from 100 $40\mu m \times 40\mu m$ RBC ROIs from our tuning data set and used hue-saturation-intensity thresholds (hue ≥ 0.95/1, saturation ≥ 0.72/1, and intensity ≥ 0.6/1). After morphological dilation with a disk-shaped structuring element of radius = $4\mu m$ (approximate radius of human red blood cells), we obtained an RBC mask and subtracted it from the nucleus map to eliminate false RBCs [28].

3.4.3.2 Lumen mapping

Lumen is typically nearly white on microscopy images. We therefore thresholded luminal pixels with values of red ≥ 0.86/1, green ≥ 0.71/1, and blue ≥ 0.82/1, using the same approach as described above (cumulative histogram based on the tuning data set).

3.4.4 Tuning ROI size and down-sampling ratio

Experimenting with our tuning data set, we selected a nearest-neighbor down-sampling ratio of 0.25, and an ROI size of $480\mu m \times 480\mu m$ ($960\ pixels \times 960\ pixels$). We ranked cancer detection performance according to the area under the receiver operating characteristic curve using FisherC with all features in a leave-one-patient-out cross-validation scheme to select these parameters [28]. These parameters were unchanged for all experiments in this chapter.

3.4.5 Feature extraction and selection

24 first-order and 132 second-order statistical features were calculated from the TCM of each ROI, giving a total of 156 features. The second-order statistical features were calculated from the grey-level co-occurrence matrix (GLCM) [31] and grey-level run length matrix (GLRLM) [32]. GLCMs and GLRLMs were calculated using neighbors in four directions without aggregation ([(0,1) represents direction 1 in the Appendix A, (-1,1) represents 2, (-1,0) represents 3, and (-1,-1) represents 4]). 22 different GLCM (neighbor distance = 1) and 11 GLRLM features were calculated. We calculated a total of 156 features; (22 GLCM + 11 GLRLM) × 4 directions + 24 first-order features = 156.

For cancer vs. non-cancer classification, the 14 top-ranked features were selected from the calculated feature set of 156 features using backward feature selection on the tuning dataset, which selects the features by ranking the AUCs from the LOPO CVs using a Fisher linear classifier. The chosen texture features are listed in the Appendix A. For high- vs. low-grade cancer classification, we selected the 41 top ranked features by using the same feature selection method for cancer vs. non-cancer classification with the tuning dataset of high-(G4) and low-(G3) grade cancer samples. The chosen features were used for both G4 vs. G3, and G4 & G5 vs. G3 grading experiments. The chosen texture features are listed in the Appendix B.

3.4.6 Cancer detection and grading using machine learning

For prostate cancer detection, we classified each ROI as cancerous vs. non-cancerous using the 14 selected features calculated from the 3-class TCMs. We performed supervised machine learning using (1) a Fisher's least square linear discriminant classifier, (2) a logistic linear classifier, and (3) a NU-SVM with a radial

basis function kernel (parameters tuned as cost = 12.5, and gamma = 0.50625 using our tuning dataset). Each of these approaches is denoted as follows throughout this chapter: (1) FisherC, (2) LoglC, and (3) SVM, respectively.

For cancer grading, we classified each ROI of all relevant cancerous regions as high- vs. low-grade by two experiments: (1) all cancerous regions of G4 and G3 for high-(G4) vs. low-(G3) classification; (2) all cancerous regions of G4 & G5 and G3 for high-(G4 & G5) vs. low-(G3) grading); using the 44 selected features calculated from the 3-class TCMs. Similarly as for cancer detection, we performed supervised machine learning using (1) FisherC, (2) LoglC, (3) SVM (a C-SVC with a linear kernel with parameters tuned as cost = 2.7, and gamma = 0.03375 by tuning using high-(G4) and low-(G3) grade samples from the tuning data set).

For both cancer detection and grading, we also used transfer learning via fine-tuning AlexNet with our nuclei maps, lumen maps, 3-class TCMs and raw image ROIs, denoted as: AlexNet-NucleiMap, AlexNet-LumenMap, AlexNet-TCMs, and AlexNet-RawIM, respectively. AlexNet was trained using 1.2 million non-medical images from the ImageNet LSVRC-2010 challenge [29]. The final fully connected layer of AlexNet was replaced by a fully connected layer with a 2-way output followed by the 2-way softmax algorithm with a 2-class label output for each of our experiments ((1) cancerous vs. non-cancerous for cancer detection, (2) high- (G4) vs. low-grade (G3), and (3) high-(G4 & G5) vs. low-(G3) grade for cancer grading). We used cross-entropy as our loss function. We used random numbers to initialize the weights and biases of the replaced layers. For all other layers, we set the initial learning rate to $\alpha = 0.0001$, and $\alpha=0.002$ for the output layer. For gradient descent, we used the adaptive moment estimation ('Adam')

optimizer [33], with β_1 = 0.9, β_2= 0.999, and $\varepsilon = 10^{-8}$ [33]. We used our tuning data set to set mini-batch size = 200, maximum epoch = 10.

We used nuclei maps, lumen maps, and 3-class TCMs as input images to fine-tune pre-trained AlexNet respectively for each of the 3 experiments. These maps were converted into RGB colour images, such that the nuclei, stroma/ other tissue, and lumen are labeled in red, green, and blue respectively. We resized each ROI with size of 240×240×3 to 227×227×3 to conform to the standard input size for AlexNet by using bilinear interpolation. We repeated the experiment using the same method with the "raw" unmodified H&E images instead of the TCMs.

3.4.7 Experiments and validation

3.4.7.1 Prostate cancer detection

We performed LOPO CV for each of the tested machine learning approaches. No same-patient samples were used in both the training and testing sets in any CV iteration. Our data set contains many more non-cancerous than cancerous ROIs in our data set. Consequently, during training, we performed random subsampling of the negative samples to balance the positive (cancerous) and negative (non-cancerous) samples. During testing, all tissue on all slides was classified. That is, we performed testing on all ROIs covering each WSI in our 68-patient set. The receiver operating characteristic (ROC) curve was computed using the cumulative predicted confidences from each trained system, and we calculated the AUC from the ROC. We calculated the cumulative error rate, FPR, and FNR by comparing the predicted labels (using the fixed operating point corresponding to the confidence level of 0.5 in all experiments) of each ROI to the designated ROI label based on the pathologist's annotations, with an ROI considered

positive when assessed by the pathologist to contain ≥ 50% cancer. The sample sizes for each tissue type are shown in Table 3.2. We also calculated the error rates (FNRs for cancerous tissue types; FPRs for non-cancerous tissue types) for each tissue type using LOPO CV.

3.4.7.2 Prostate cancer grading

We performed the same CV as for cancer detection, for each approach for high- vs. low-grade cancer classification. During training, we balanced the positive (high-grade) and negative (low-grade) samples by random duplication of the samples for whichever class had the smaller sample size. The validation was done by comparison to the pathologist's annotations, with an ROI considered high-grade when assessed by the pathologist to contain ≥ 50% 1) G4 for high-(G4) vs. low-(G3) grading, and 2) G4 or G5 for high-(G4 & G5) vs. low-(G3) grading. For high-(G4 & G5) vs. low-(G3) grading, we calculated the FNRs for the high-grade cancer tissue types using LOPO CV.

3.5 References

1. Stephenson, A.J., et al., *Predicting the outcome of salvage radiation therapy for recurrent prostate cancer after radical prostatectomy.* J Clin Oncol, 2007. **25**(15): p. 2035-41.

2. Izawa, J.I., *Salvage radiotherapy after radical prostatectomy.* Can Urol Assoc J., 2009. **3**(3): p. 245-250.

3. Gleason, D.F., G.T. Mellinger, and G. Veterans Administration Cooperative Urological Research, *Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. 1974.* J Urol, 2002. **167**(2 Pt 2): p. 953-8; discussion 959.

4. Epstein, J.I., et al., *The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System.* Am J Surg Pathol, 2016. **40**(2): p. 244-52.

5. Epstein, J.I., et al., *A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score.* Eur Urol, 2016. **69**(3): p. 428-35.

6. Kryvenko, O.N. and J.I. Epstein, *Prostate Cancer Grading: A Decade After the 2005 Modified Gleason Grading System.* Arch Pathol Lab Med, 2016. **140**(10): p. 1140-52.

7. Evans, A.J., et al., *Interobserver variability between expert urologic pathologists for extraprostatic extension and surgical margin status in radical prostatectomy specimens.* Am J Surg Pathol, 2008. **32**(10): p. 1503-12.

8. Epstein, J.I., et al., *The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma.* Am J Surg Pathol, 2005. **29**(9): p. 1228-42.

9. Sun, M., et al., *Insights of modern pathology reports originating from prostate biopsy and radical prostatectomy specimens.* Eur Urol, 2012. **62**(1): p. 40-1.

10. van der Kwast, T.H., et al., *International Society of Urological Pathology (ISUP) Consensus Conference on Handling and Staging of Radical Prostatectomy Specimens. Working group 2: T2 substaging and prostate cancer volume.* Mod Pathol, 2011. **24**(1): p. 16-25.

11. Montironi, R., et al., *Handling of radical prostatectomy specimens: total embedding with large-format histology.* International journal of breast cancer, 2012. **2012**.

12. Croke, J., et al., *Proposal of a post-prostatectomy clinical target volume based on pre-operative MRI: volumetric and dosimetric comparison to the RTOG guidelines.* Radiat Oncol, 2014. **9**: p. 303.

13. Gibson, E., et al., *Registration of prostate histology images to ex vivo MR images via strand-shaped fiducials.* J Magn Reson Imaging, 2012. **36**(6): p. 1402-12.

14. Soetemans, D.J., *Computer-assisted characterization of prostate cancer on magnetic resonance imaging.* Electronic Thesis and Dissertation Repository, 2017. **4504**.

15. Mosquera-Lopez, C., et al., *Computer-Aided Prostate Cancer Diagnosis From Digitized Histopathology: A Review on Texture-Based Systems.* IEEE Rev Biomed Eng, 2015. **8**: p. 98-113.

16. Doyle, S., et al., *A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies.* IEEE Trans Biomed Eng, 2012. **59**(5): p. 1205-18.

17. Litjens, G., et al., *Automated detection of prostate cancer in digitized whole-slide images of H and E-stained biopsy specimens*. SPIE Medical Imaging. Vol. 9420. 2015: SPIE.

18. Litjens, G., et al., *Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis.* Sci Rep, 2016. **6**: p. 26286.

19. Monaco, J.P., et al., *High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models.* Med Image Anal, 2010. **14**(4): p. 617-29.

20. DiFranco, M.D., et al., *Ensemble based system for whole-slide prostate cancer probability mapping using color texture features.* Comput Med Imaging Graph, 2011. **35**(7-8): p. 629-45.

21. Nguyen, K., A.K. Jain, and B. Sabata, *Prostate cancer detection: Fusion of cytological and textural features.* J Pathol Inform, 2011. **2**: p. S3.

22. Rashid, S., et al., *Automatic pathology of prostate cancer in whole mount slides incorporating individual gland classification.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2019. **7**(3): p. 336-347.

23. Nir, G., et al., *Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts.* Med Image Anal, 2018. **50**: p. 167-180.

24. Leo, P., et al., *Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images.* J Med Imaging (Bellingham), 2016. **3**(4): p. 047502.

25. Boyce, B.F., *Whole slide imaging: uses and limitations for surgical pathology and teaching.* Biotech Histochem, 2015. **90**(5): p. 321-30.

26. Kwak, J.T. and S.M. Hewitt, *Nuclear architecture analysis of prostate cancer via convolutional neural networks.* IEEE Access, 2017. **5**: p. 18526-18533.

27. Gorelick, L., et al., *Prostate histopathology: learning tissue component histograms for cancer detection and classification.* IEEE Trans Med Imaging, 2013. **32**(10): p. 1804-18.

28. Han, W., et al., *Automatic cancer detection and localization on prostatectomy histopathology images*. SPIE Medical Imaging. Vol. 10581. 2018: SPIE.

29. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

30. Ruifrok, A.C. and D.A. Johnston, *Quantification of histochemical staining by color deconvolution.* Anal Quant Cytol Histol, 2001. **23**(4): p. 291-9.

31. Haralick, R.M., K. Shanmugam, and I. Dinstein, *Textural Features for Image Classification.* IEEE Transactions on Systems, Man, and Cybernetics, 1973. **SMC-3**(6): p. 610-621.

32. Galloway, M.M., *Texture analysis using grey level run lengths.* Computerized Medical Imaging and Graphics, 1975. **4**: p. 172-179.

33. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

Chapter 4

A version of this chapter is currently in preparation for journal publication: Wenchao Han, Michelle Downes, Theodorus van der Kwast, Joseph Chin, Stephen Pautler, and Aaron Ward, "Automatic cancer subtype grading on digital histopathology images of radical prostatectomy specimens."

4 Automatic cancer subtype grading on digital histopathology images of radical prostatectomy specimens

4.1 Introduction:

Prostate cancer (PCa) has been the most commonly diagnosed non-skin cancer in men in Canada since 1998 [1]. It is a highly variable disease, with some cases being indolent and others lethal. Early detection and treatment play important roles in curing the disease and improving survival. Radical prostatectomy (RP) has been demonstrated to have excellent disease control [2] but many patients still suffer from recurrence and metastasis [3]. Patient risk management is important for selecting and guiding treatment post-surgery (e.g. adjuvant/salvage therapy). The Gleason grading system was first proposed by Donald Gleason et al [4] in 1966. It stratifies tumours into five aggressiveness levels based solely on their morphological patterns that have been correlated to clinical outcomes. The Gleason score is a summation of the Gleason grades of the primary and secondary tumours (based on tumour size and corresponding Gleason grade). The Gleason score at RP is considered to be the most powerful indicator for predicting disease progression [5, 6] and therefore is an essential factor for patient stratification.

Clinical reporting of Gleason grades has the intention of optimally reflecting tumour progression; this is supported by clinical studies correlating Gleason grades with patients' clinical outcomes. In the most recently updated scoring system, Gleason grade 3 (G3) usually involves small and infiltrative glands, with large variations in the amount of intervening stroma to form a more sparse or dense tumour. G4 includes the subtypes of cribriform glands (including the glomeruloid pattern), poorly formed glands, and fused glands [7]. G5 includes the subtypes of solid nests, cords of cells, individual cells, or nests of cribriform glands with unequivocal necrosis [7]. In the latest scoring system, cribriform patterns are assigned as G4 instead of G3 since they correlate with poorer clinical outcomes. In addition, the newly proposed grade groups separate G3+4, G4+3, and G4+4 into three grade groups based on the amount of G4 presenting within the tumour, since studies have demonstrated different prognostic outcomes from these three grade groups [8].

There is now increasing interest in finding pathological indicators beyond the Gleason score [9, 10] for informing interventions leading to better disease control and treatment planning. Dong et al. [11] found that the presence of cribriform G4 was an independent predictor of biochemical recurrence and metastasis after surgery. In addition, they found that the presence of poorly formed glands, fused glands, and cribriform glands simultaneously indicated shorter biochemical disease-free survival. Therefore, they suggest the potential need for sub-classification of G4 patterns beyond the Gleason grading system. Trudel et al. [12] found that any amount of cribriform G4/intraductal carcinoma was a significant predictor of biochemical recurrence-free rate after adjusting for Gleason score and T stage. They suggested the reporting of those features

independently upon validation using other clinical outcomes as end-points, with a larger cohort of data.

From these studies and the findings of recent ISUP meetings [6, 8], we may surmise that the presence and amounts of sub-Gleason morphological patterns could be informative for predicting disease progression. Since there are many sub-patterns within each Gleason grade, further studies may be needed to find other indicators having different prognostic values. Also, to incorporate such findings into routine clinical practice, larger studies are generally needed to support the value of reporting those features clinically. Conducting such studies is technically difficult since they require manual annotation and subtyping of a large number of RP sections for each individual tumour. Subtype reporting is not a standard clinical procedure and requires extra time from busy pathologists, challenging the practicality of such studies. Therefore, there is a need for an automatic tool which automatically annotates cancerous region of RP specimens according to Gleason subtype. Such a tool would enable the automated annotation of large numbers of tissue sections, making correlation of subtype with clinical outcomes more straightforward and practical.

Many published methods have demonstrated the potential for computer-assisted detection and grading of PCa on digital histopathology images. A recent review summarizes texture-based approaches to these problems [13], with more recent publications showing very high accuracies using deep learning-based approaches for PCa detection [14]. However, there is no study in the literature demonstrating the use of an automated tool for PCa subtype grading. PCa subtype grading is a more challenging problem since the levels of differentiation of the cancer cells are more similar within each

Gleason grade than across different Gleason grades and thus are visually more similar in terms of tissue patterns and harder for the learning algorithm to distinguish.

In our previous work, we developed and validated automated systems, which 1) detect and map cancerous regions on whole-slide-images (WSIs) of whole-mount RP sections with an area under the receiver-operating-characteristic curve (AUC) of 0.98 [15] and 2) grade cancer as high vs. low grade with an AUC of 0.92 [16]. In this work, our objective is to develop and validate a system which grades cancerous regions of interest (ROIs) according to eight subtypes (i.e. sparse G3, packed G3, intermediate G3, desmoplastic G3, large cribriform G4, small fused G4, poorly formed G4, benign intervening; see Figure 4.1 for examples).

4.2 Materials and Methods

4.2.1 Materials

This study was approved by our institution's Human Subjects Research Ethics Board with informed consent of all patients. From 25 RP patients, we obtained 92 mid-gland whole slide images (WSIs) of whole-mount tissue sections (tissue sizes are approximately $3 \times 4\ cm^2$). The tissue sections were cut at $4\mu m$ from paraffin blocks and stained with hematoxylin and eosin (H&E). We scanned using a ScanScope (Leica Biosystems, Wetzlar, Germany) at 20X ($0.5\mu m/pixel$) in bigTiff format with no compression.

4.2.2 Methods

4.2.2.1 Method overview

Each WSI was annotated (via contouring and sub-grading of each tumour region) by two genitourinary pathologists (Downes and van der Kwast). The subtypes were indicated by contour colours (see the example WSIs in Figure 4.4 and Figure 4.5). Our implementation used Matlab 2018b (The Mathworks, Natick, MA).

Our method consists of five steps, described in detail as follows.

4.2.2.2 Tumour annotation and subtype grading

Each of the RP sections was reviewed cooperatively by two pathologists (van der Kwast and Downes with 28 years and 11 years of experience, respectively). They identified tumour foci and assigned a Gleason grade to each. They further evaluated G3 and G4 tissues to assign a subtype. G3 was divided into desmoplastic, sparse, intermediate, and packed G3 (Figure 4.1); the latter three subtypes were divided based on the proportion of the intervening stroma tissue and the apposition of glands. The G3 patterns have separate glands which do not merge with their neighbours. They were separated by various amounts of stroma tissue. Sparse G3 contains predominantly stroma tissue with scattered glands, while packed G3 contains dense glands with minimal intervening stroma tissue. G4 was separated into 1) large cribriform, 2) small fused glands, 3) poorly formed glands (Figure 4.1). The annotation was performed on digitized whole-mount WSI using different colours to represent different subtypes. All annotations were performed at 2× by agreement of the two pathologists. After the annotation, a review was performed at higher power [17].

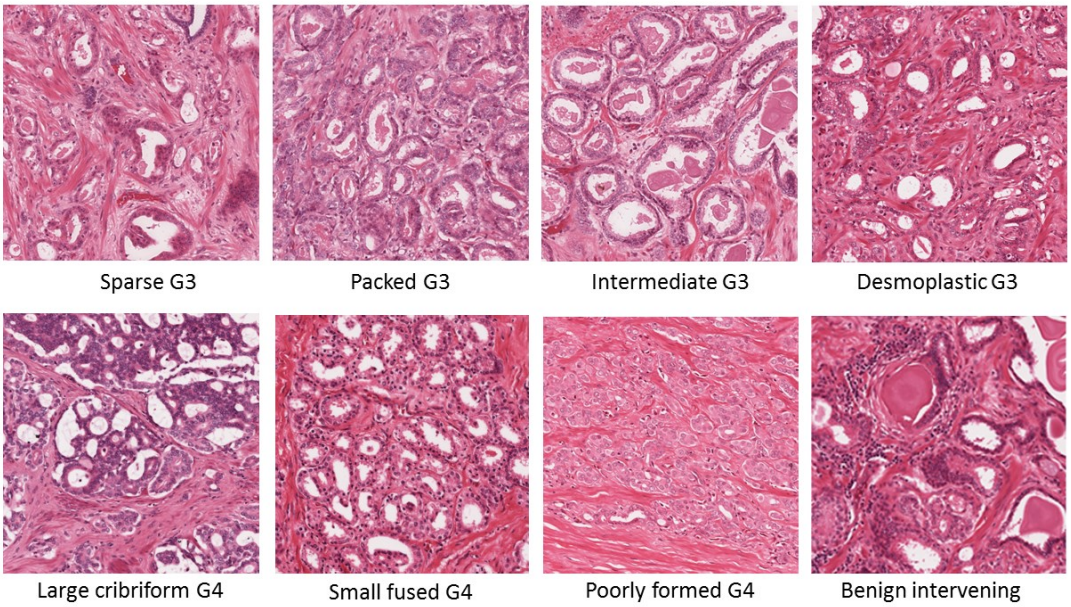


Figure 4.1: 480 × 480 µm samples of each of the tissue types classified in this chapter. Note the heterogeneity of appearance of the tissues within each Gleason grade. Note also the similarity of appearance of tissues across different Gleason grades in some cases (e.g. packed G3 vs. small fused G4).

4.2.2.3 ROI labeling

Each WSI was separated into a set of square 480×480 µm ROIs, with the ROI size chosen by experimenting ROI sizes from 120×120 µm to 600×600 µm with 120 µm incremental intervals on our tuning data set for cancer vs. non-cancer classification [15]. ROIs containing more than 50% cancerous tissue were used in our experiment and a subtype label was assigned to each ROI according to the manual pathology annotations. For an ROI containing a mixture of more than two subtypes of cancer tissue, we assigned the subtype label using the subtype with the largest amount of tissue in the ROI.

4.2.2.4 System training

For each subtype, we fine-tuned AlexNet [18] (pre-trained using the ImageNet [19] data set) by training the network using our labeled ROI samples. We down-sampled each ROI to 227×227×3 pixels to conform to the necessary input size for AlexNet. We replaced the final AlexNet layer with a fully connected layer, which has a 2-way output followed by 2-way softmax with a two possible outputs: the subtype on which the network was trained, or other (i.e. the output label indicates whether or not the ROI contains the subtype on which the classifier was trained). We calculated the loss function using cross-entropy. The weights and biases of the replaced layers were initialized with random numbers. We set the initial learning rate to be $\alpha = 0.0001$ for all the other layers, and $\alpha = 0.002$ for the output layer to make the weights and biases from other layers almost unchanged while those from the output layer learn faster. We used the adaptive moment estimation ("Adam") optimizer [20] for gradient descent. The hyper-parameters were set as: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$ based on [20]. Other training parameters were set as: mini-batch size = 200, maximum epoch = 10. We chose all the hyper-parameters in accordance with our previous study on cancer grading [16].

4.2.2.5 Classification

We performed supervised machine learning for each experiment classifying: (1) sparse G3, (2) packed G3, (3) desmoplastic G3, (4) intermediate G3, (5) large cribriform G4, (6) poorly formed gland G4, (7) small fused G4, (8) benign intervening vs. negative samples (i.e. all the samples which were not defined as positive in the experiment) respectively. During training, the positive and negative samples were balanced by replication of positive samples.

4.2.2.6 Validation

We performed leave-one-WSI-out cross-validation (CV) using all relevant ROIs throughout all WSIs with testing on all WSIs containing both positive and negative samples for each fold. We calculated the AUC using the classifier's confidence value and the assigned ROI labels (see ROI labeling) according to the pathologists' annotations. We calculated the false positive rate (FPR) and the false negative rate (FNR) using the classifier confidence threshold which yielded the closest operating point to top left corner of the receiver-operating-characteristic (ROC) curve. We also calculated the error rate using the same operating point.

Table 4.1: Sample sizes (number of $\boldsymbol{480\mu m \times 480\mu m}$ ROIs) for each subtype.

G3	Sparse G3	1867
	Intermediate G3	4140
	Packed G3	893
	Desmoplastic G3	515
G4	Large cribriform G4	275
	Small fused G4	2423
	Poorly formed G4	1099
Benign	Benign intervening	225

4.3 Results

Table 4.1 shows the sample size of each subtype used in the experiments. Table 4.2 shows the error metrics for cancer subtype grading, classifying each ROI either as the given subtype in the row, or not the subtype in the given row. Since we have a small sample size for some subtypes and not all WSIs had all the subtypes, we computed

cumulative error metrics per-ROI from the leave-one-WSI-out CV. The system yielded AUCs larger than 0.7 for all the subtypes except for packed G3. The subtypes of desmoplastic G3, small fused G4, poorly formed G4, and benign intervening have AUCs larger than 0.8.

Table 4.2: Error metrics from leave-one-WSI-out cross-validation classifying each subtype.

	Error Rate (%)	FNR (%)	FPR (%)	AUC
Sparse G3	26.8	29.9	26.2	0.78
Intermediate G3	36.0	32.0	38.3	0.70
Packed G3	50.2	34.0	51.5	0.58
Desmoplastic G3	22.0	26.8	21.8	0.82
Large cribriform G4	23.7	33.8	23.4	0.74
Small fused G4	23.1	28.6	21.5	0.82
Poorly formed G4	23.8	18.4	24.4	0.86
Benign intervening	26.6	20.9	26.7	0.82

Figure 4.2 shows the FPR for each subtype for each of the eight experiments. For example, for the classifier trained to detect sparse G3, the FPR of 33.8% for the negative sample of intermediate G3 means that 33.8% of the intermediate G3 samples were falsely labeled as sparse G3. We found that the FPRs were higher for classifiers trained to detect intermediate G3, packed G3 and poorly formed G4, in comparison to the other classifiers. For the classifier trained to detect packed G3, most of the FPRs were larger than 40% and two of them (i.e. desmoplastic G3 and benign tissue) were close to 40%. For the classifier trained to detect intermediate G3, we found FPRs larger than 50% for the negative samples of sparse G3 and packed G3, which were followed closely by large cribriform

G4 and poorly formed G4. We found substantially lower FPRs for the negative samples of the other subtypes. For the classifier trained to detect poorly formed G4, we found higher FPRs for the negative samples of packed G3, desmoplastic G3, large cribriform G4, and benign intervening. The highest FPR was found for the negative samples of large cribriform G4.

Since this is the first preliminary study to attempt the challenging task of classifying Gleason subtypes using machine learning, our intention is to generate observations and hypotheses that could be helpful to further engineering efforts to render a robust tool that will be useful in research studies automatically correlating the presence of Gleason subtypes with ultimate clinical outcomes. To that end, Figure 4.3 shows scatter plots correlating the amount of each subtype observed in each patient by the pathologists, versus the amount predicted by the classifiers. Each data point on the scatter plot represents an individual patient. We found that all the classifiers had sensitivities of 100% (i.e. no data points lie at zero on the horizontal axis and at non-zero points on the vertical axis). However, this was at the expense of specificity as indicated by the points on the scatter plots showing cases where the system overestimated the amounts of the subtypes. This is also illustrated by the points lying at zero on the vertical axis and at non-zero points on the horizontal axis; these are false positive detections. In general, the false positives involve smaller amounts of detected cancer, compared to the true positives. The system yielded higher specificities for the G3 subtypes (except for desmoplastic G3) compared to the G4 subtypes. We found the systems performed best for sparse G3 and intermediate G3, which have the fewest false positives and better

agreements between manual annotated tumour size and system predicated tumour size, compared to the other subtypes.

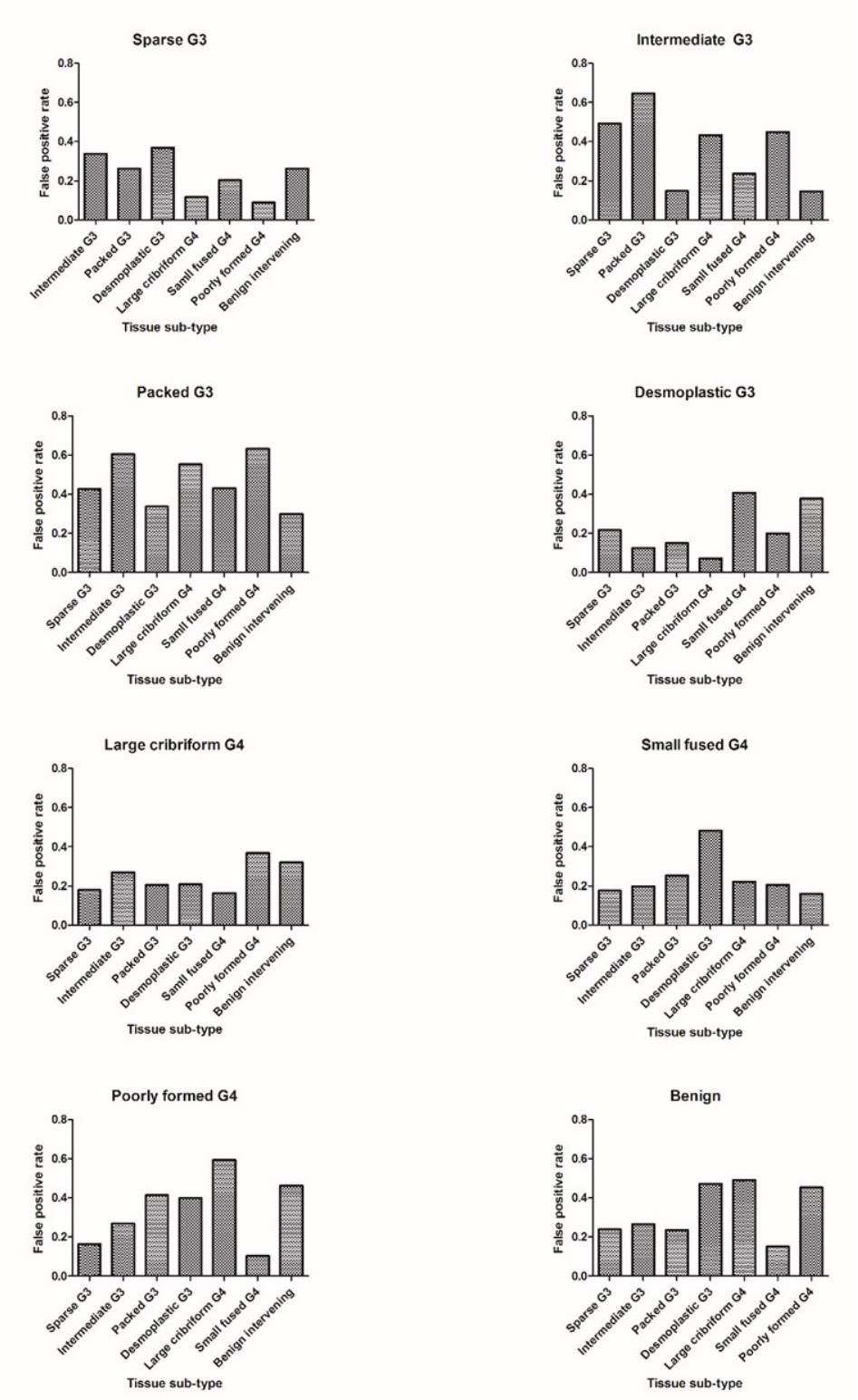


Figure 4.2: FPRs for each tissue subtype, broken down by confounding subtype.

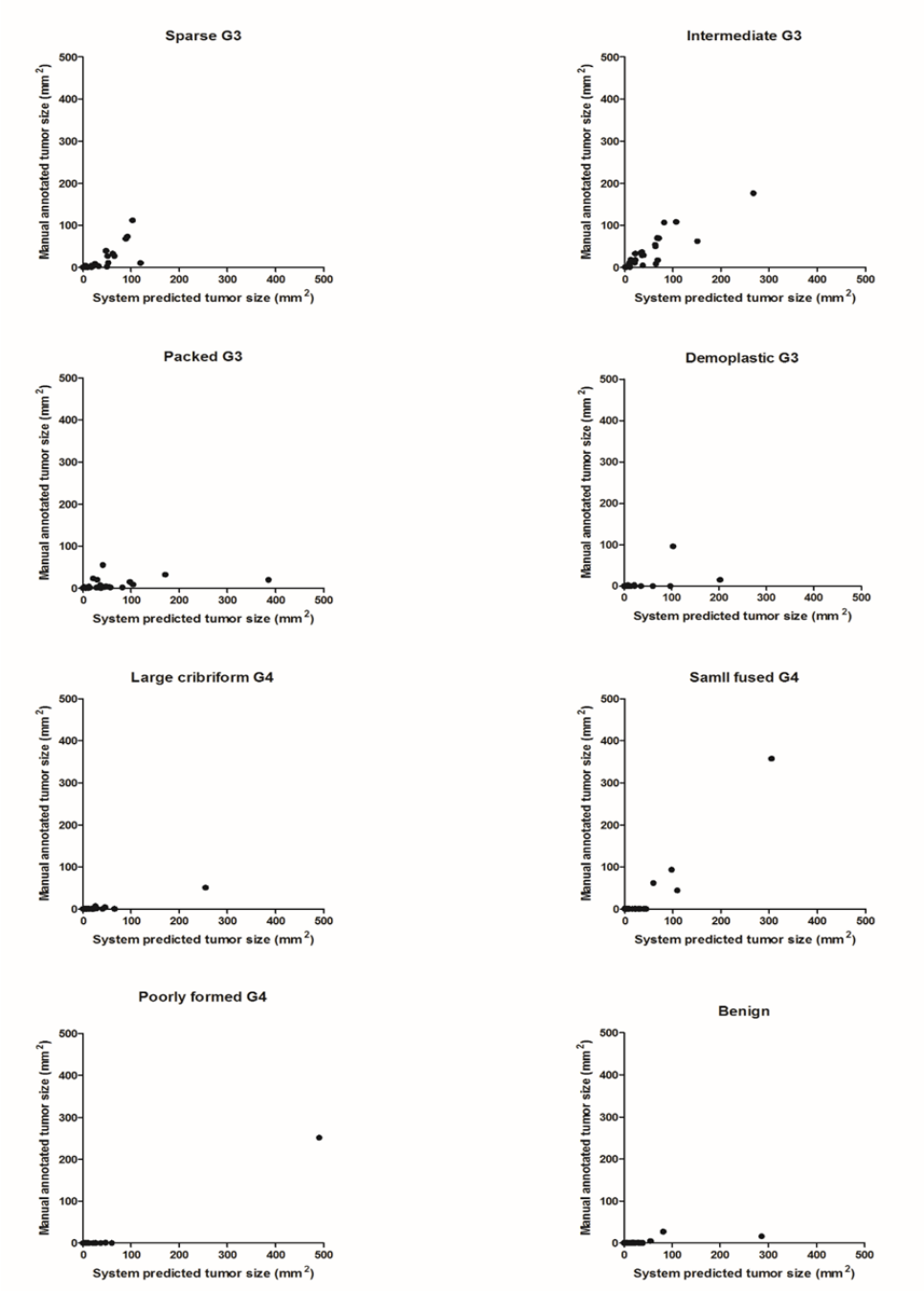


Figure 4.3: Scatter plot of manual annotated tumour size per patient vs. system predicted tumour size per patient for each subtype

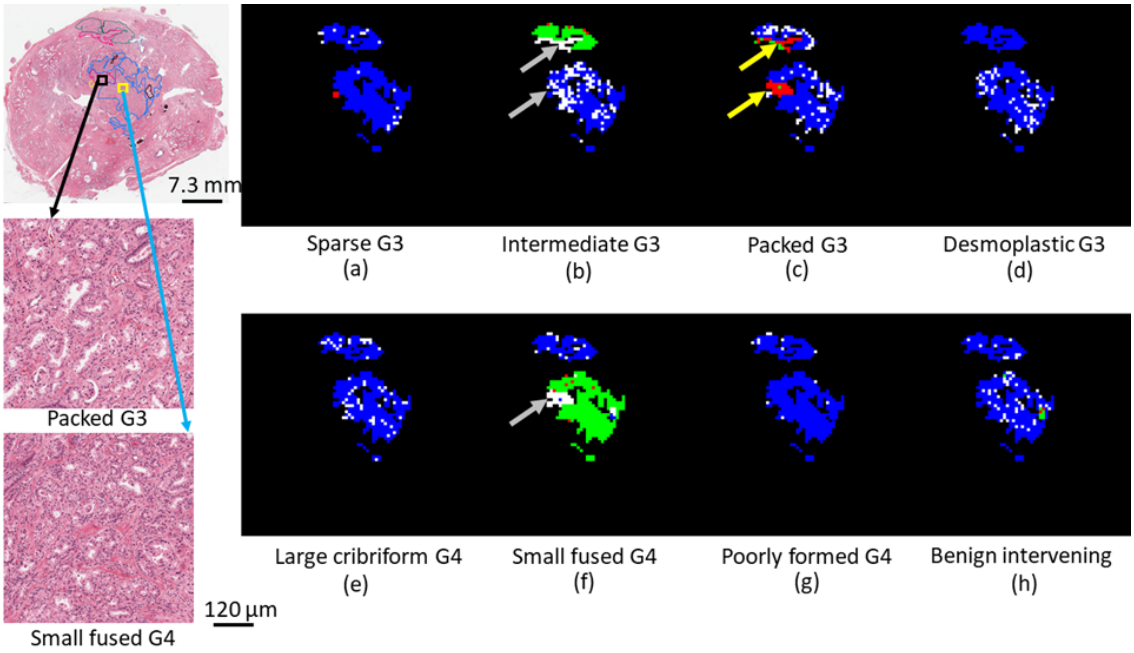


Figure 4.4: Label maps from an example WSI for PCa sub-grading. Left column: example WSI with two tissue samples shown below zoomed in from the black and yellow boxes on the WSI. (a) – (h) are label maps after validating system predicated results against manual annotation. Map annotations: Blue = true negative, Green = true positive, Red = false negative, white = false positive. Pathologist's annotations on histology: Pink = packed G3, blue = small fused G4, dark green = intermediate G3, brown = benign intervening. Yellow arrows indicate the false negative regions. Grey arrows indicate the false positive regions.

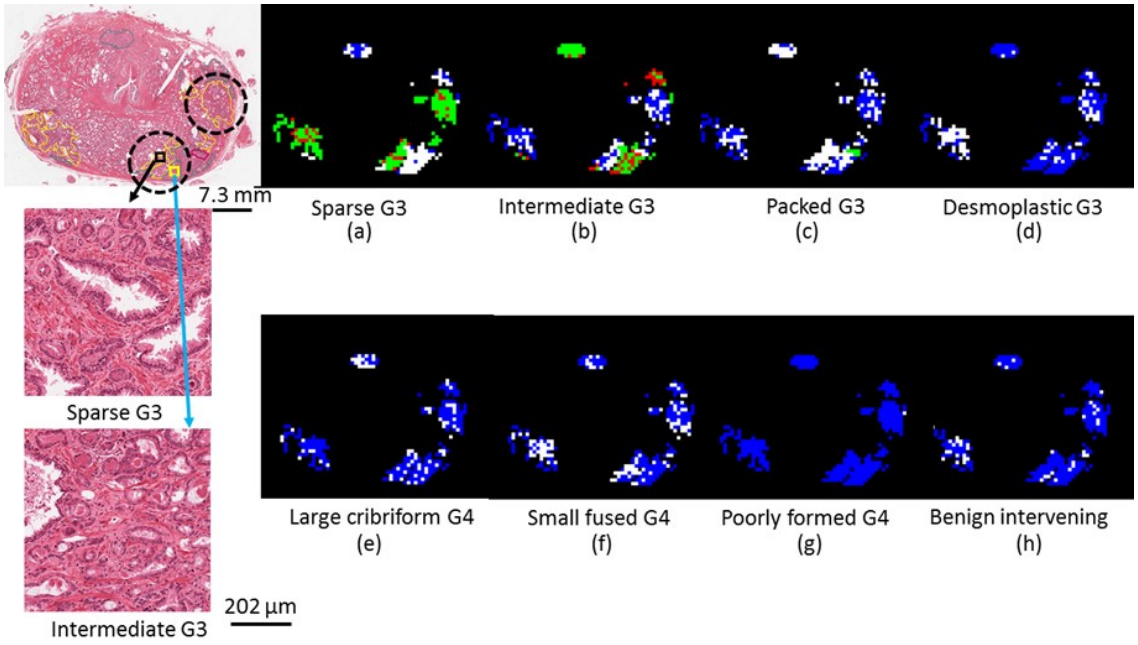


Figure 4.5: Label maps from an example WSI for PCa sub-grading. Left column: example WSI with two tissue samples shown below zoomed in from the black and yellow boxes on the WSI. (a) – (h) are label maps after validating system predicated results against manual annotation. Map annotations: Blue = true negative, Green = true positive, Red = false negative, white = false positive. Pathologist's annotations on histology: Pink = packed G3, yellow = Sparse G3, dark green = intermediate G3. Black circle: highlighted regions.

Figure 4.4 and Figure 4.5 show the systems' outputs for two selected cases. The case in Figure 4.4 contains confounding cancer regions of packed G3 and small fused G4. The case in Figure 4.5 contains confounding subtypes of sparse, intermediate, and packed G3. Our system outputs were validated against our pathologists' annotations in the form of label maps for detecting the subtype regions from each of the eight experiments. For the first case (see Figure 4.4), we found that generally the eight predicted label maps correctly mapped out most of the positive and negative regions. The major errors happened in the regions of packed G3, which were missed by the system and falsely labeled as small fused G4 and intermediate G3 (see the yellow and grey arrows pointed regions in Figure 4.4 (b), (c), and (f)). We also observe that those regions of packed G3 with false detections appeared adjacent to cancerous regions containing small fused G4 and intermediate G3 to form tumours, instead of appearing in the form of independent tumours. For the second case (see Figure 4.5), similarly, generally most system outputs correctly labeled positive and negative regions. The major errors were within the G3 tumour regions. The black circled regions show the conflicting results from the systems for detecting sparse G3 and intermediate G3 (Figure 4.5 (a) and (b)), with both systems labeling the region as positive. We found that the system trained to detect packed G3 had an overall oversensitivity in terms of labeling both regions of sparse G3 and intermediate G3 as positive (Figure 4.5 (c)).

4.4 Discussion

In general, our experiments demonstrated promising results for this novel and challenging problem, using binary classification. The systems yielded AUCs higher than 0.7 for detecting all of the subtypes, except for the system for detecting packed G3. Four

of the eight systems yielded AUCs higher than 0.8 (Table 4.2). FPRs and FNRs were generally smaller than 35% for most of the systems (Table 4.2), which suggests the systems were sensitive and specific in detecting those subtypes. For most of the systems, the FPRs for each subtype were generally lower than 40%, with most of them close to 20% (see the bar charts for each subtypes in Figure 4.2). This suggests that the systems are generally capable of differentiating the subtypes from each other, except for a few confounding subtypes. On the qualitative results we illustrated, the systems successfully labeled the primary tumour and secondary tumour as small fused G4 and intermediate G3 (Figure 4.4 (b) and (f)) respectively, with good agreement among each of the systems (Figure 4.4). We can see that, for the primary tumour, all the systems labeled the major region as negative except for the system trained to detect packed G3 (Figure 4.4 (c)). Similarly, the systems labeled most of the tumour regions correctly (Figure 4.5).

Although, in principle, larger sample sizes should improve the systems' performance, some subtypes may be harder to differentiate than others and vice versa. We found that the systems yielded good performance even for some subtypes with relatively smaller sample sizes. For detecting subtypes of desmoplastic G3 and benign intervening, the system yielded AUCs of 0.82; they have sample sizes of 515 and 225, respectively. Similarly, for large cribriform G4, although we have a small sample size of 275, the system still yielded an AUC of 0.74 (see Table 4.1 and Table 4.2). We speculate that this is because those subtype tissue appearances are more unique in morphology, which is not confounding with respect to other subtypes. However, for some subtypes, even if a large sample size was given, the systems still had inferior performance. This can be evidenced by the relatively poorer performance in detecting subtypes of intermediate

G3 and packed G3. For intermediate G3, although the largest sample size of 4140 was given, which was 10 to 20 times than that of the subtypes mentioned above, the system yielded lower AUC of 0.7. In addition, for packed G3, with a sample size of 893, we found the lowest AUC of 0.58 (see Table 4.1 and Table 4.2).

We observed that the G3 subtypes are confounding to each other. This may be because they have similar morphological patterns. For intermediate G3, we found that the highest FPRs (> 50%) were for sparse and packed G3 (see plot of intermediate G3 in Figure 4.2). Likewise, for packed G3, we found high FPRs for sparse and intermediate G3 (see plot of packed G3 in Figure 4.2). Although, for sparse G3, the FPRs were generally much lower, the two highest FPRs were for the subtypes of intermediate, packed, and desmoplastic G3 (see plot of sparse G3 in Figure 4.2). From the graphical results of the example cases (Figure 4.4 and Figure 4.5), we observed false detections among the sparse G3 (Figure 4.5 (a)), intermediate G3 (Figure 4.5 (b)), and packed G3 (Figure 4.5 (c)), and the over sensitivity of the system trained to detect packed G3 falsely labeling regions of different subtype tissues as positive (Figure 4.5 (c)). These results make sense because those three subtypes of G3 (i.e. sparse, intermediate, and packed G3) were similar in morphology; all of them have similar gland structure with different amounts of intervening stroma tissue (see the two example ROIs of sparse G3 and intermediate G3 in Figure 4.5; similar gland structure was observed with different amounts of intervening stroma). Also, since our method classified non-overlapped ROIs with a fixed ROI size, sampling error may negatively affect system performance. For example, a sampled ROI could capture a region which containing primarily stroma tissue, with limited glandular tissue from the region containing intermediate G3. This can result

in the system incorrectly labeling the ROI as sparse G3. In the process of manual annotation, the pathologist usually first identifies the ROI and then zooms in and out to estimate the region to assign the subtype. In this way, a better estimation of stroma proportion may be achieved.

We speculate that the varied nature of spatial patterns of packed G3 makes correctly detecting packed G3 difficult (see AUC for packed G3 in Table 4.2). For packed G3, the system tended to be oversensitive to most of the subtype tissues except for desmoplastic G3 and benign tissue, which is evidenced by overall high FPRs for most subtypes (see plot of packed G3 in Figure 4.2). We speculate this is because the gland structure of packed G3 is similar to that of sparse and intermediate G3 and it was so dense that the system trained to detect packed G3 was trained to recognize those gland structures to label them as positive for any ROI containing similar gland structures in sparse and intermediate G3 regions (see example ROI of packed G3 in Figure 4.4, and sparse and intermediate G3 in Figure 4.5). In addition, we speculate that since the cancer tissue containing packed G3 may be in the process of progression to G4, they often appear simultaneously and intertwined with each other. This means that samples of packed G3 may unavoidably contain some intervening G4 patterns and thus the system was trained to recognize those patterns as packed G3 (see the example ROI of packed G3 in Figure 4.4, which contains small portions of fused glands).

Although, for detecting most of the subtypes, the system yielded generally good performance and differentiation capability for each of the subtypes, there are certain subtypes that are more confounding than others. We found that the desmoplastic G3 and small fused gland G4 were confounding to each other (see the bar charts of desmoplastic

G3 and small fused G4 in Figure 4.2). Poorly formed G4 and large cribriform G4 are confounding to each other (see the bar charts of poorly formed G4 and large cribriform G4 in Figure 4.2). Systems may be specifically developed for differentiating those confounding subtypes from each other (e.g. developing a system for classifying desmoplastic G3 vs. small fused G4).

It is important to consider the results at the per-patient level, in addition to the overall AUC across all ROIs. This is because 1) the system performance at the patient level may more directly reflect the potential for application to studies correlating subtype presence and amount with outcome; 2) overall AUC results do not fully reflect the system performance at the patient level for subtype detection (i.e. to find the presence of any amount of the subtype for the patient). The systems performed better for subtypes having ROIs across larger numbers of patients. For example, for the sparse and intermediate G3, they have higher specificity and better agreement of manual estimated tumour size and system predicted tumour sizes than the other subtypes (Figure 4.3). However, their AUCs were not the highest (Table 4.2). They have relatively larger sample sizes of ROI across more patients (Table 4.1 and Figure 4.3). Also, we found that the system yielded higher specificity for the G3 subtypes than for the G4 subtypes (Figure 4.3) although some of the G4 subtypes (i.e. poorly formed G4 and small fused G4) have higher AUCs than the G3 subtypes (Table 4.2). However, better agreement of tumour size estimation was found for the true positives of those G4 subtypes (Figure 4.3), which have relatively large cancer areas.

At the patient level, for the purpose of subtype detection (i.e. to find the presence of any amount of the subtype for the patient), for all the subtypes, we found (Figure 4.3)

that 1) the systems' sensitivities were 100%, 2) the systems' sensitivities were higher than their specificities, 3) the systems' false positives were smaller than their true positives, especially for the G4 subtypes. This suggests that the systems are overly sensitive for most of the subtypes, and small positive regions are less likely to be true positives, compared to large positive regions. Therefore, it may be helpful in future work to consider post-processing the system outputs by thresholding based on the reported tumour size for some subtypes (e.g. cribriform G4, poorly formed G4, and small fused G4). Future work on performance improvements at the patient level should focus primarily on the G4 subtypes, with a larger sample size at both of the ROI and patient level.

Our results should be interpreted in the context of several limitations. First, since we used annotations determined by consensus of two pathologists, inter-observer variability was not taken into consideration. Also, all of our tissues were processed and stained in a single laboratory; although the samples were acquired over several years and we observed substantial staining variability due to different batches of stain and other factors. Thus, validation using samples from multiple laboratories with annotations by different observers is essential to translation of this tool to use in medical research. Second, although there is evidence that using transfer learning can train a deep network more efficiently (i.e. using smaller sample sizes) [21], some subtypes with small samples may not be fully representative of the variability seen in the wider population, potentially explaining the inferior results we observed for some subtypes. Using data augmentation and/or a larger sample size may mitigate this issue. Third, our results of system performance at the patient level should be interpreted with limitations of using small

sample sizes for most of the subtypes. Fourth, we only used one type of deep learning method; there are many deep learning methods are available that demonstrate good performance in different image identification tasks. Further exploration using those methods may improve the system performance.

To the best of our knowledge, this work represents the first validation of a machine learning based system for PCa subtype grading on whole-mount surgical pathology specimens where validation was conducted throughout all of the relevant tissues on each entire slide, such that ROIs from the same WSI were never used for both training and testing in any fold. The system achieved AUCs higher than 0.8 for subtypes of desmoplastic G3, small fused G4, poorly formed G4, and benign intervening; AUCs higher than 0.7 for subtypes of sparse G3, intermediate G3, and large cribriform G4.

In conclusion, we demonstrated that transfer learning by fine-tuning AlexNet can classify G3 and G4 PCa into eight subtypes on WSIs of H&E stained RP tissue sections based on the Gleason grading system, validated against pathologist annotations using leave-one-WSI-out CV. We found that the systems generally yield promising performance for potential clinical use/studies for detecting most of the subtypes upon further technical improvement and validation on large data set. The subtype of packed G3 was challenging to detect due to the varied nature of its spatial patterns. The intermediate G3 subtype was confounding to other subtypes of G3, which may because they share many gland features and were differentiated by the proportion of stroma tissue only. However, the system yielded higher specificity for detection at the patient level for the G3 subtypes than for the G4 subtypes. Our future work includes developing and validating 1) systems for differentiating mutually confounding subtypes (e.g. packed G3

vs. intermediate G3), 2) an algorithm to combine multiple binary systems into a multiclass system for the subtype grading, and 3) a post-processing method for the subtypes to improve system specificity at the patient level.

4.5 References

1. Levy, I.G., N.A. Iscoe, and L.H. Klotz, *Prostate cancer: 1. The descriptive epidemiology in Canada.* Cmaj, 1998. **159**(5): p. 509-513.

2. Wong, Y.-N., et al., *Survival associated with treatment vs observation of localized prostate cancer in elderly men.* Jama, 2006. **296**(22): p. 2683-2693.

3. D, B. and W. T, *Adjuvant radiotherapy after radical prostatectomy: indications, results and side effects.* Urologia Internationalis, 2007.

4. Gleason, D., *Histological grading and clinical staging of prostatic carcinoma.* Urologic pathology. The prostate, 1977. **171**.

5. Roberts, W.W., et al., *Contemporary identification of patients at high risk of early prostate cancer recurrence after radical retropubic prostatectomy.* Urology, 2001. **57**(6): p. 1033-1037.

6. Epstein, J.I., et al., *The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma.* Am J Surg Pathol, 2005. **29**(9): p. 1228-42.

7. Kryvenko, O.N. and J.I. Epstein, *Prostate cancer grading: a decade after the 2005 modified Gleason grading system.* Arch Pathol Lab Med, 2016. **140**(10): p. 1140-1152.

8. Epstein, J.I., et al., *The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System.* Am J Surg Pathol, 2016. **40**(2): p. 244-52.

9. Kryvenko, O.N., et al., *Gleason score 7 adenocarcinoma of the prostate with lymph node metastases: analysis of 184 radical prostatectomy specimens.* Archives of Pathology and Laboratory Medicine, 2013. **137**(5): p. 610-617.

10. Kweldam, C.F., et al., *Cribriform growth is highly predictive for postoperative metastasis and disease-specific death in Gleason score 7 prostate cancer.* Modern pathology, 2015. **28**(3): p. 457.

11. Dong, F., et al., *Architectural heterogeneity and cribriform pattern predict adverse clinical outcome for Gleason grade 4 prostatic adenocarcinoma.* Am J Surg Pathol, 2013. **37**(12): p. 1855-1861.

12. Trudel, D., et al., *Prognostic impact of intraductal carcinoma and large cribriform carcinoma architecture after prostatectomy in a contemporary cohort.* Eur J Cancer, 2014. **50**(9): p. 1610-1616.

13. Mosquera-Lopez, C., et al., *Computer-Aided Prostate Cancer Diagnosis From Digitized Histopathology: A Review on Texture-Based Systems.* IEEE Rev Biomed Eng, 2015. **8**: p. 98-113.

14. Litjens, G., et al., *Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis.* Sci Rep, 2016. **6**: p. 26286.

15. Han, W., et al., *Automatic cancer detection and localization on prostatectomy histopathology images*. SPIE Medical Imaging. Vol. 10581. 2018: SPIE.

16. Han, W., et al., *Automatic high-grade cancer detection on prostatectomy histopathology images*. SPIE Medical Imaging. Vol. 10956. 2019: SPIE.

17. Downes, M.R., et al., *Determination of the association between T2-weighted MRI and Gleason sub-pattern: a proof of principle study.* Academic radiology, 2016. **23**(11): p. 1412-1421.

18. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

19. Deng, J., et al. *Imagenet: A large-scale hierarchical image database*. in *2009 IEEE conference on computer vision and pattern recognition*. 2009. Ieee.

20. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

21. Shin, H.C., et al., *Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning.* IEEE Trans Med Imaging, 2016. **35**(5): p. 1285-98.

Chapter 5

5 Conclusions and future directions

5.1 Contributions

This thesis addresses the unmet need for automatic PCa detection, grading, and subtype grading on H&E stained WSIs of whole-mount RP tissue sections via technological advancements and validation experiments. Methods were developed for processing WSIs of whole-mount RP sections for automatic tissue component segmentation, and tissue region classification. Validation was conducted on 286 expert annotated whole-mount WSIs from 68 patients, which cover all clinically relevant grade groups, using CV with data grouped on a per-patient basis. A subset of patients were used for subtype grading. Comparisons between conventional machine learning based approaches and deep learning based approaches were performed. Also, comparisons using different inputs between raw images and different TCMs were performed. In addition, a system for automatic PCa subtype grading was first proposed and validated. Thus, this thesis contributes to a better understanding of how machine learning approaches work and the utility of each method for PCa detection, grading and subtype grading. The accomplishment of the objectives (Chapter 1 section 1.3) and answering the research questions (Chapter 1 section 1.3) led to the following advancements in technology and knowledge.

5.1.1 Advances in knowledge and technology arising from this thesis

A tissue component segmentation algorithm was proposed to segment 3-class tissue components (i.e. nuclei, lumen, stroma/other) from H&E stained histology

images efficiently, to compensate for staining variability for accurate and robust PCa detection and grading on whole-mount WSIs of RP sections.

Chapter 2 presented a tissue component segmentation method to segment the tissue image into nuclei, lumen, and stroma/other to allow for efficient feature extraction for accurate and efficient classification of cancer vs. non-cancer (Chapter 2) and high- vs. low-grade (Chapter 3) cancer image samples across different WSIs from different patients. Staining variability negatively affects classification results by creating inconsistency in the extracted features [1]. Previously work [2-4] used normalization algorithms to reduce the staining variation. These methods used selected images for calibration, or normalized images to each other. The biggest limitation of these methods is that they were dependent on the pre-selected target images for calibration, or images used in the experiment for normalization. Some methods were proposed [5, 6] to segment tissue components from the tissue images with or without normalization using machine learning based approaches. Similarly, these methods were dependent on the image samples used for training. This dependency leads to the limitations in the generalization capability of the proposed systems and therefore in the capability for clinical translation. To create a consistent feature set that is robust to staining variability, we proposed a segmentation method for fast and accurate tissue component segmentation to generate the 3-class TCM. This allows our machine learning systems to use features extracted from those computed 3-class TCMs to achieve the highest AUCs of 0.96 for PCa detection (Chapter 2) and 0.89 for PCa grading (Chapter 3). It also allows us to use transfer learning using 3-class TCMs as input to achieve AUCs of 0.98 for PCa detection (Chapter 2) and 0.93 for PCa grading (Chapter 3). The key component in this

segmentation method is the proposed adaptive thresholding algorithm, which yields fast and accurate nuclei segmentation despite large staining variability. This algorithm is independent of other images and fast in computation. *Therefore, this thesis contributes to tissue component segmentation on H&E-stained whole-mount WSIs by providing a novel algorithm for fast and accurate nuclei segmentation that can compensate for staining variability.*

A machine learning pipeline using conventional machine learning approaches demonstrated the feasibility and efficiency of using machine learning to analyze whole-mount WSIs for PCa detection and grading for clinical translation.

Chapters 2 and 3 presented a machine learning pipeline for automatic PCa detection and grading on WSIs of whole-mount RP sections. PCa detection and grading on WSIs of whole-mount RP sections provide valuable information for post-surgery patient care, which can be life-saving by advising adjuvant therapy for appropriate patients. Also, annotating and grading each individual tumour on whole-mount RP sections can potentially support clinical studies for better patient stratification post-RP. Manual annotation is time consuming and infeasible in the standard clinical workflow and requires a substantial amount of extra effort from pathologists. Therefore, there is an unmet need for a system that can automatically annotate and grade each tumour on whole-mount WSIs of RP sections. Previous work primarily presented methods using pre-selected ROIs [7], which has limitations in scalability and generalization capability to be applied to whole-mount WSIs due to the substantially large amount of data from the whole-mount WSIs and the heterogeneity of PCa tissues. Two studies [6, 8] presented methods for PCa detection using whole-mount WSIs but are limited in the ability to

detect high-grade cancer tissue, which is of high prognostic value. No study was found for PCa grading using all cancerous tissue that includes all clinically relevant grade groups from whole-mount RP sections. Our proposed pipeline demonstrated state-of-the-art performance for both PCa detection and grading with 25 minutes/whole-mount WSI. Our systems were validated on 286 WSIs from 68 patients with data grouped on a per-patient basis. This was the first study for PCa detection and grading using all tissues from whole-mount WSIs, which covers all clinically relevant grade groups. *Therefore, this thesis contributes to the technological advancement in automatic PCa detection and grading on whole-mount WSIs using machine learning by developing and validating a machine learning pipeline yielding state-of-the-art performance with practical processing time.*

A deep learning pipeline was presented for PCa detection, grading and subtype grading using transfer learning with pre-trained AlexNet, which yields superior performance, compared to the conventional machine learning approaches. The deep learning pipeline also allowed comparisons among different machine learning methods and comparisons among the uses of different TCMs as inputs in the context of our research objectives (Chapter 1 section 1.3).

Chapters 2, 3 and 4 presented a transfer learning approach by fine-tuning pre-trained AlexNet for automatic PCa detection, grading and subtype grading on WSIs of whole-mount RP sections. The purpose is similar to that discussed above. In addition, subtypes of cribriform G4 and intraductal carcinoma were found to have independent prognostic value in a few studies [9-12], thus further studies/trials on this topic may support reporting those parameters in the clinical routine and discovering new biomarkers

for patient care. Previous work has demonstrated excellent performance using deep learning approaches to analyze digital histopathology images for PCa detection on TMAs and WSIs of biopsies [13-16]. One study used a deep learning approach for PCa grading on TMAs and reported an overall accuracy of 77.8%. Our work was the first study using deep learning detecting and grading PCa on whole-mount WSIs. Also, our study demonstrated the first automatic system for PCa subtype grading beyond the Gleason grading system. In previous work, most studies [13, 15, 16] used raw image samples as system input with minimal pre-processing. One study [14] used nuclei maps as input and demonstrated superior performance in comparison to using raw image as input. We used raw images, 3-class TCMs, nuclei maps, and lumen maps as system inputs and found that 3-class TCMs yielded the best overall performance and the second best performance for identifying high-grade cancer for both PCa detection and grading. *Therefore, this thesis demonstrated a technological advancement in automated systems for PCa detection, grading and subtype grading on WSIs of whole-mount RP sections using deep learning.*

5.1.2 Answers to central research questions:

1. Can features extracted from TCMs provide the major information for PCa detection and grading on whole-mount RP sections?

For PCa detection, using first- and second-order statistical features extracted from the 3-class TCMs, all the three classifiers (i.e. a Fisher linear discriminant classifier, a logistic linear classifier, and SVM) achieved AUCs $\geq$ 0.92 with highest AUC of 0.96 using SVM (Chapter 2). For grading, all classifiers achieved AUCs $\geq$ 0.82 with highest AUC of 0.89 using Fisher classifier (Chapter 3). These results demonstrated state-of-the-art performance in comparison to the studies using first- and second- order statistical

features extracting directly from the raw images without object segmentations [7]. Also, our results outperformed the results in the previous study [5] using a subset of our data set with features extracted from 9-class TCMs. Although the comparisons were made with some limitations (e.g. each study used different validation methods and data sets), generally our results, which were validated on the largest data set with data grouped on a per-patient basis, demonstrated that the 3-class TCMs provided major information for using the first- and second- order statistical features for PCa detection and grading.

In addition, transfer learning fine-tuned with 3-class TCMs achieved the best performance with AUCs of 0.98 and 0.92 for PCa detection (Chapter 2) and grading (Chapter 3). In comparison, transfer learning using raw images yielded AUCs of 0.98 and 0.92 for PCa detection and grading, respectively. The almost identical (difference is at the third decimal) performance using raw images as input suggested that 3-class TCMs provided equivalent information to the raw images for PCa detection and grading using pre-trained AlexNet. Although we only used one type of classifier for the direct comparison in our study, the reported results demonstrated state-of-the-art performance in comparison to the literature in which a few other deep learning approaches were used with raw images as inputs [13, 15, 16]. *Therefore, features extracted from the 3-class TCMs can provide major cues for PCa detection and grading.*

2. Can 3-class TCMs compensate for staining variability for robust PCa detection and grading on whole-mount RP specimens?

Staining variability is a big issue, which negatively affects the performance using computational methods to analyze histopathology images [1]. Therefore, previous studies

used features extracted directly from the raw images after image normalization or calibration, or features extracted at the object-level after image segmentation to resolve this issue. However, previous methods' normalization and calibration approaches render them specific to the data set at hand, limiting their potential for generalization. Also, the validations were not grouped on a per-patient basis in most studies, such that image samples stained differently might be used for system training or normalization to overfit the data. Thus, the generalization capability of those methods is difficult to evaluate. In our study, large staining variability among WSIs from different patients was observed qualitatively (Chapter 2). Despite the staining variability, our systems yielded state-of-the-art performance for PCa detection and grading using the resulting 3-class TCMs (Chapters 2 and 3). The validations were conducted using data grouped on a per-patient basis such that the systems were validated against the staining variability between the WSIs from different patients. The computation of the 3-class TCMs is independent of other images. *Thus, the thesis suggested that the 3-class TCMs can compensate for the staining variability for robust PCa detection and grading.*

3. What is the most important information on the histology tissue that can be used for PCa detection and grading?

Chapters 2 and 3 compared using TCMs to using raw images as input for PCa detection and grading. Chapter 2 demonstrated that 3-class TCMs provided the major information for PCa detection, and Chapter 3 extended that knowledge to PCa grading. One previous work [14] found that a nuclei seed map (i.e. maps reflecting nuclei location information only) is essential for PCa detection by demonstrating that systems using nuclei seed maps yielded the best performance comparing to using raw image as input.

Our work compared the performance using raw images, 3-class TCMs, nuclei maps, and lumen maps for both PCa detection and grading. We found that, for both PCa detection and grading, using 3-class TCMs yielded the best performance followed closely by using raw images and nuclei maps, and using nuclei maps yielded the best performance for identifying G5 involved tissue (i.e. G5, G4+5, G5+4). *Thus, this thesis advances the knowledge that nuclei maps encode the primary information of 3-class TCMs for PCa detection and grading and are essential for identifying G5 involved tissue.*

4. How do deep learning based approaches perform for PCa detection and grading?

Although previous studies [13-16] demonstrated excellent performance using deep learning methods to approach those problems, direct comparisons to those using conventional machine learning methods is difficult since the data sets and validation methods were different between different studies. A few studies [14, 15] provided more direct comparisons between methods for PCa detection and grading. However, the comparisons were purely conducted at the overall performance level using relatively small sample sizes which were not close to the scale that literature reported for achieving good performance [17, 18]. Our studies performed direct comparisons between the deep learning and conventional machine learning methods using the 3-class TCMs for both PCa detection and grading at the overall performance level and the level for each tissue type (Chapters 2 and 3). We found that pre-trained AlexNet outperformed conventional machine learning based approaches for both PCa detection (AUC of 0.98 vs. 0.96) and grading (AUC of 0.92 vs. 0.89) in overall performance. Similar performance was achieved for tissue types having small (i.e. sample size smaller than 1000) or medium sample sizes (i.e. sample size larger than 1000 and smaller than 4000). Pre-trained

AlexNet outperformed conventional machine learning methods for the tissue types having large sample sizes (i.e. sample size $\geqslant$ 4000). *Therefore, pre-trained AlexNet yielded better performance than the conventional machine learning methods for our problems when using 3-class TCMs as inputs.*

In addition, comparison between AlexNet fine-tuned by raw images and AlexNet fine-tuned by 3-class TCMs in the training sample size test (Chapter 2) and in the performance at the tissue type level (Chapters 2 and 3) demonstrated that although raw images provide more information to the system, AlexNet fine-tuned with raw images is more sensitive to training sample size than with the 3-class TCMs. Reducing the complexity of the input image (i.e. using 3-class TCMs) may yield more robust performance to sample size. *Thus, this thesis advances knowledge that system performance is sensitive to sample size using pre-trained AlexNet and using high-level features as input can reduce this sensitivity.*

5. What is the feasibility of detecting the subtypes of G3 and G4 PCa using a deep learning approach?

Chapter 4 validated a deep learning approach for detecting each of eight G3 and G4 PCa subtypes on whole-mount WSIs using leave-one-WSI-out cross-validation. There is no previous work on this topic. The cumulative results computed across all ROIs measured the system's capability for identifying each subtype from the others. The FPR breakdown for each subtype further quantifies the system's capability for differentiating each subtype from each of the other subtypes. In general, the results are promising for subtype detection, as our overall cumulative AUCs are larger than or equal to 0.7 for

seven of the eight subtypes and the AUCs are larger than or equal to 0.8 for four of the eight subtypes. For most of the eight experiments, the most FPRs for each subtype were close to 20%. This demonstrated that the system is capable of differentiating each subtype from each of the other subtype most of the time. However, some subtypes are harder to detect (e.g. intermediate, packed G3) and this might be because they are confounding to most of the G3 subtypes, and packed G3 was also confounding to some G4 subtypes. These speculations were supported by the relatively higher FPRs among those confounding subtypes (e.g. for detecting packed G3, FPRs ≥ 40% for sparse G3, intermediate G3, large cribriform G4, small fused G4 and poorly formed G4).

The performance of using this method for clinical research needs to be further validated on multi-centre studies with a large data set. For subtypes of sparse and intermediate G3, the system performance at the patient level is better than the other subtypes. The performance evaluation at the patient level (Chapter 4) provided a more direct evaluation which reflects the system application for potential clinical studies. We found that the cumulative AUCs did not reflect the systems' capability of detecting the presence of any amount of the subtype at the patient level. The system showed the best performance for sparse and intermediate G3 at the patient level. Although the system yielded high AUCs for the G4 subtypes, the specificities were much lower than those for the G3 subtypes. We found that the subtypes of sparse and intermediate G3 have large sample sizes of ROIs across more patients while, for the G4 subtypes, generally most ROIs were concentrated in a few patients. Therefore, further studies with larger sample size at the ROI and patient levels are needed to further evaluate the feasibility of the system at the patient level for clinical research translation. *Thus, this thesis advances the*

knowledge of the feasibility and challenges of using machine learning to detect PCa subtypes beyond Gleason grade.

5.2 Limitations

The content of this thesis should be interpreted in the context of several limitations of our studies. All tissues used in our studies were processed in the same clinical pathology laboratory using manual staining. Since tissue processing conditions and protocols vary from centre to centre, our studies were limited by not taking the variability from multiple centres into account. We would expect this to affect the methods using raw images more than those using TCMs, the computation of which is adaptive and calibration-free. Second, the cancerous annotations were done by one physician and verified by one of two pathologists in Chapters 2 and 3. Although two pathologists reviewed the cases used in Chapter 4, we used annotations from their consensus. These aspects of our studies limit the variability of the observers' contours. Third, there are many classifiers for deep learning and conventional machine learning approaches. Our results and conclusions are limited to the classifiers we used in our studies. Fourth, the adaptive thresholding algorithm was performed at the WSI level, which makes the computed threshold a global threshold for each WSI. Our informal experiments (not reported here) suggested that adaptive thresholding on sub-regions of each WSI did not improve performance. However, adapting our method to do if needed is straightforward. Finally, it must be acknowledged that all CV studies may be subject to positive bias in their results; therefore, validation using an external data set is required to support clinical translation of this tool.

5.3 Applications and future directions

5.3.1 Discovering biomarkers beyond Gleason grading system for clinical patient care

Although the Gleason grading system has shown its prognostic value, is widely accepted and used clinically, there may be morphological patterns or pattern groups beyond the Gleason grading system, which are better associated with prognostic outcome. The Gleason grading system was revised multiple times since first proposed due to new discoveries in clinical practice and studies/trials [19]. In the 2014 ISUP consensus meeting [20], the Gleason grading system was updated. Some pathological patterns were reassigned into different Gleason grades, and the newly developed grade group system was presented, which showed better correlation to prognostic outcomes than the old grade group system [20]. Also, some pathology patterns were recommended to be reported independently such as small cell carcinoma, which has an unique appearance with poorer prognostic outcome than poorly differentiated tumours. In addition, some subtypes (i.e. cribriform G4 and intraductal carcinoma) were found to have independent prognostic value in recent studies [11, 12]. Therefore, revisions to the existing pathology reporting may be continued with new findings to further revise the Gleason grading/grade group systems, or even to be extended beyond the existing system by proposing new systems.

Because prostate tissue is heterogeneous, there are many patterns beyond Gleason grade [21]. Some patterns may be independently prognostic or different combinations of patterns may be prognostic. Methods developed in this thesis demonstrated the capability of differentiating morphological patterns of histopathology prostate tissues based on the

Gleason grading system (Chapters 2 and 3) and beyond the Gleason grading system (Chapter 4). Our methods have the potential to be used to identify different morphological patterns and relate those to clinical outcomes directly, to evaluate the need for reporting those parameters in routine pathological reports. This may be achieved by using an unsupervised machine learning approach [22, 23], which is a machine learning approach using unlabeled samples to classify those samples into groups, with patient outcome as an endpoint. In Chapter 2, 156 features were calculated from TCMs. Although there may be some redundancy in those features, they are expected to provide the capability of differentiating patterns beyond Gleason grade. In addition, there are many other features [e.g. scale-invariant feature transform (SIFT) feature set [24]] that can be calculated for differentiating image patterns. Thus, those features could be used with unsupervised machine learning to identify prognostic patterns beyond Gleason grades. Because superior performance using deep learning was found in this thesis, and raw images can be used as input, fine-tuning a deep network with raw images can potentially be used to identify patterns using information beyond the 3-class TCMs via combining this with unsupervised machine learning.

5.3.2 Translational applications for other disease/tissue types

The machine learning pipeline developed in this thesis has the potential to be adapted to detect other diseases on WSIs of whole-mount sections. Our work focused on PCa detection and grading while other studies focused on cancer detection/grading on histopathology images of other organ tissues such as breast [25], oropharyngeal [26], colorectal [27], and brain [28] tissue. Since our methods rely on identifying morphological patterns of H&E stained tissue samples, which are typically used for other

organs [29], our pipeline is transferable to be applied for detecting and grading diseases in other organs. This translation may be beneficial due to the advantages of being capable of processing whole-mount WSIs, efficiency in terms of processing time, and robustness to staining variability, which were demonstrated in our studies.

The segmentation method presented in this thesis may be extended to be applied to other H&E-stained tissues to support quantitative histomorphometry [29] or analyzing tumour morphology and its most invasive elements. Many studies [29] have proposed methods for nuclei segmentation from different perspectives (e.g. using different methods, on different tissue, or for different application purposes). Our algorithm for nuclei segmentation relies on the phenomenon that generally the nuclei and other tissue components have different amounts of hematoxylin stain. This phenomenon applies to many other organ tissues such as breast, colorectal, and brain tissues. Therefore, our algorithm has the potential to segment nuclei for other organ tissues, with the advantages of being robust to staining variability, amenability to be implementation on WSIs of large tissue sections, and efficiency in terms of computation.

5.3.3 Remaining gaps in knowledge toward clinical translation and future directions

Although there is an unmet need for using computational tools to support more efficient and easier routine clinical pathology reporting as previously discussed (Chapter 1 section 1.3) and there are numerous publications [29] on developing computer assisted systems for pathology reporting, there are still gaps to translate those tools into routine clinical practice. One of the major knowledge gaps toward translational application is the capability of conducting comprehensive validation studies. Since most studies used pre-

selected ROIs, validations using all tissue samples with a large data set is required to fully validate the system performance. Although the systems proposed in this thesis were validated on all tissue of WSIs of whole-mount RP sections, we did not validate on data sets from different centres and did not take inter-observer variability into consideration. Nir et al. [15] have approached those questions using pre-selected samples. Multi-centre studies taking all those factors into consideration using a large data set without tissue preselection, which covers enough variability in staining, tissue patterns, observers and artifacts, will support clinical translation.

In addition, the error metrics used in this thesis and other studies may not appropriately evaluate the efficacy of automated systems in routine clinical application. We may envision that the role of those systems is assisting pathologists to examine the tissue to generate pathology reports. Although good performance (e.g. high AUC, low error rate, high recall etc.) was reported in this thesis, those error metrics may not reflect efficacy in clinical practice. Relating those error metrics to clinical performance of those systems is essential. For example, a system reporting very high error rate of 40% with 100% recall may be practical since pathologists may only need to examine system-labeled positives. This may be approached by large-scale user studies. For example, since a system may yield different performances by choosing different operating points on the ROC curve, a proper threshold may be chosen for specific clinical use through user studies.

5.4 Reference

1. Leo, P., et al., *Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images.* J Med Imaging (Bellingham), 2016. **3**(4): p. 047502.

2. Magee, D., et al. *Colour normalisation in digital histopathology images*. in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*. 2009. Daniel Elson.

3. Basavanhally, A. and A. Madabhushi, *EM-based segmentation-driven color standardization of digitized histopathology*. SPIE Medical Imaging. Vol. 8676. 2013: SPIE.

4. Mosquera-Lopez, C. and S. Agaian, *Iterative local color normalization using fuzzy image clustering*. SPIE Defense, Security, and Sensing. Vol. 8755. 2013: SPIE.

5. Gorelick, L., et al., *Prostate histopathology: learning tissue component histograms for cancer detection and classification.* IEEE Trans Med Imaging, 2013. **32**(10): p. 1804-18.

6. Rashid, S., et al., *Automatic pathology of prostate cancer in whole mount slides incorporating individual gland classification.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2019. **7**(3): p. 336-347.

7. Mosquera-Lopez, C., et al., *Computer-Aided Prostate Cancer Diagnosis From Digitized Histopathology: A Review on Texture-Based Systems.* IEEE Rev Biomed Eng, 2015. **8**: p. 98-113.

8. Monaco, J.P., et al., *High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models.* Med Image Anal, 2010. **14**(4): p. 617-29.

9. Kryvenko, O.N., et al., *Gleason score 7 adenocarcinoma of the prostate with lymph node metastases: analysis of 184 radical prostatectomy specimens.* Archives of Pathology and Laboratory Medicine, 2013. **137**(5): p. 610-617.

10. Kweldam, C.F., et al., *Cribriform growth is highly predictive for postoperative metastasis and disease-specific death in Gleason score 7 prostate cancer.* Modern pathology, 2015. **28**(3): p. 457.

11. Dong, F., et al., *Architectural heterogeneity and cribriform pattern predict adverse clinical outcome for Gleason grade 4 prostatic adenocarcinoma.* Am J Surg Pathol, 2013. **37**(12): p. 1855-1861.

12. Trudel, D., et al., *Prognostic impact of intraductal carcinoma and large cribriform carcinoma architecture after prostatectomy in a contemporary cohort.* Eur J Cancer, 2014. **50**(9): p. 1610-1616.

13. Litjens, G., et al., *Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis.* Sci Rep, 2016. **6**: p. 26286.

14. Kwak, J.T. and S.M. Hewitt, *Nuclear architecture analysis of prostate cancer via convolutional neural networks.* IEEE Access, 2017. **5**: p. 18526-18533.

15. Nir, G., et al., *Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts.* Med Image Anal, 2018. **50**: p. 167-180.

16. Chen, P., et al., *An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis.* Nature medicine, 2019. **25**(9): p. 1453-1457.

17. Krizhevsky, A., I. Sutskever, and G.E. Hinton. *Imagenet classification with deep convolutional neural networks*. in *Advances in neural information processing systems*. 2012.

18. Shin, H.C., et al., *Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning.* IEEE Trans Med Imaging, 2016. **35**(5): p. 1285-98.

19. Kryvenko, O.N. and J.I. Epstein, *Prostate cancer grading: a decade after the 2005 modified Gleason grading system.* Arch Pathol Lab Med, 2016. **140**(10): p. 1140-1152.

20. Epstein, J.I., et al., *The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System.* Am J Surg Pathol, 2016. **40**(2): p. 244-52.

21. Downes, M.R., et al., *Determination of the association between T2-weighted MRI and Gleason sub-pattern: a proof of principle study.* Academic radiology, 2016. **23**(11): p. 1412-1421.

22. Nattkemper, T.W. and A. Wismüller, *Tumor feature visualization with unsupervised learning.* Med Image Anal, 2005. **9**(4): p. 344-351.

23. Iglesias-Rozas, J. and N. Hopf, *Histological heterogeneity of human glioblastomas investigated with an unsupervised neural network (SOM).* Histology and histopathology, 2005.

24. Lowe, D.G., *Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image*. 2004, Google Patents.

25. Basavanhally, A., et al., *Multi-field-of-view framework for distinguishing tumor grade in ER+ breast cancer from entire histopathology slides.* IEEE transactions on biomedical engineering, 2013. **60**(8): p. 2089-2099.

26. Lewis Jr, J., et al. *A Quantitative Histomorphometric Classifier Identifies Aggressive Versus Indolent p16 Positive Oropharyngeal Squamous Cell*

Carcinoma. in *LABORATORY INVESTIGATION*. 2013. NATURE PUBLISHING GROUP 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA.

27. Bychkov, D., et al., *Deep learning based tissue analysis predicts outcome in colorectal cancer.* Sci Rep, 2018. **8**(1): p. 3395.

28. Ertosun, M.G. and D.L. Rubin. *Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks*. in *AMIA Annual Symposium Proceedings*. 2015. American Medical Informatics Association.

29. Madabhushi, A. and G. Lee, *Image analysis and machine learning in digital pathology: Challenges and opportunities*. 2016, Elsevier.

Supplementary material for Chapter 3

Appendix A: 14 selected features for cancer vs. non-cancer classification. GLCM: grey level co-occurrence matrix. GLRLM: grey level run length matrix. IDM: inverse difference moment. IMC: information measure of correlation. 1, 2, 3, 4: one of the 4 directional offsets used for calculating the matrix.

Mean gradient value	GLCM entropy-2
GLCM IDM-2	GLCM correlation-1
GLCM IDM-3	GLCM cluster shade-3
GLRLM short run emphasis-3	GLCM IMC2-1
GLRLM short run low gray level emphasis-1	GLCM IMC2-2
GLRLM short run low gray level emphasis-3	GLCM energy-1
GLRLM short run low gray level emphasis-4	GLCM energy-2

Appendix B: selected features for high- vs. low-grade cancer classification. GLCM: grey level co-occurrence matrix. GLRLM: grey level run length matrix. IDM: inverse difference moment. IMC: information measure of correlation. 1, 2, 3, 4: one of the 4 directional offsets used for calculating the matrix.

Gray level bimodality coefficient	GLCM difference entropy-1
GLCM correlation-1	GLCM IMC1-3
GLCM correlation-2	GLCM IMC2-1
GLCM correlation-4	GLCM IMC2-4
GLCM variance-1	GLCM IDM-2
GLCM sum average-3	GLCM IDM-3
GLRLM short run emphasis-2	GLCM IDM-4
Proportion of stroma	GLRLM run length nonuniformity-1
Gray level variance	GLRLM run percentage-4
Gray level uniformity	GLRLM short run high gray level emphasis-2
GLCM entropy-3	GLRLM long run low gray level emphasis-1
GLCM variance-3	GLRLM long run low gray level emphasis-2
GLCM variance-4	Gray level entropy
GLCM sum average-1	GLCM cluster prominence-3
GLCM sum average-4	GLCM sum average-2
GLCM sum entropy-1	GLCM difference entropy-2
GLCM sum entropy -3	GLCM difference entropy-3
GLCM sum entropy -4	GLRLM long run emphasis-2
GLCM sum variance-1	GLRLM gray level nonuniformity-3
GLCM sum variance-3	GLRLM low gray level run emphasis-2
GLCM sum variance-4	

Permission for Reproduction of Published Materials

Juan F Madrid <jfmadrid@um.es>
Mon 2019-10-28 11:40 AM
Wenchao Han

The permission is granted by HISTOLOGY AND HISTOPATHOLOGY. The article in Histology and Histopathology should be properly cited.

Yours sincerely,

Prof. Juan F. Madrid, Editor
HISTOLOGY AND HISTOPATHOLOGY
Department of Cell Biology and Histology
School of Medicine
University of Murcia
E-30100 Espinardo - Murcia
Spain
FAX: +34-868884150

El 27/10/19 a las 19:06, Wenchao Han escribió:

Dear Professor Juan F. Madrid:

My name is Wenchao Han. I am a PhD candidate from the department of Medical Biophysics, University of Western Ontario, Canada.

I am writing my PhD thesis right now and hope to use one figure from the paper "Contemporary approaches for processing and handling of radical prostactomy specimens." I am writing this email to ask for the reprint permission from you.

The use of this reprint is for academic research only. Upon successfully defending my thesis, and the thesis passed final review, the electronic thesis will be published through scholarship@western. My thesis title will be: "Automatic cancer detection, grading, subtype grading on digital histopathology images of whole-mount radical prostatectomy sections."

The paper citation information is:
M.-T. Sung and L. Cheng, "Contemporary approaches for processing and handling of radical prostactomy specimens," Histology and histopathology, 2010.

The link of the paper is:
https://digitum.um.es/digitum/bitstream/10201/42577/1/Contemporary%20approaches%20for%20processing%20and%20handling%20of%20radical%20prostactomy%20specimens.pdf

I would like to reprint Figure 1 in page 262. The figure detail is in below:
Fig. 1. Processing of radical prostatectomy specimen.

For your convenience, I have attached the PDF version of the paper on this email for reference.

Shall you have any questions regarding this request, please do not hesitate to contact me through this email.

Thank you

Best,

Wenchao

Academic UK Non Rightslink <permissionrequest@tandf.co.uk>
Mon 2019-10-28 4:58 AM
Wenchao Han

Dear Wenchao Han

Figure 5 from C Higgins (2015) Applications and challenges of digital pathology and whole slide imaging, Biotechnic & Histochemistry, 90:5, 341-347, DOI: 10.3109/10520295.2015.1044566

Thank you for your correspondence requesting permission to reproduce the above material from our Journal in your printed thesis and to be posted in your university's repository.

We will be pleased to grant entirely free permission on the condition that you acknowledge the original source of publication and insert a reference to the Journal's web site: www.tandfonline.com

Please note that this licence does not allow you to post our content on any third party websites or repositories.

Thank you for your interest in our Journal.

Best wishes

Karin Beesley - Permissions Administrator, Journals
Taylor & Francis Group
3 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN, UK
Permissions Tel: +44 (0)20 7017 7617
Permissions e-mail: permissionrequest@tandf.co.uk

SPRINGER NATURE

GENERAL TERMS AND CONDITIONS

1. Grant of Rights

1.1. Subject to payment by the Licensee of the Licence Fee (where applicable), the Licensor grants to the Licensee the non-exclusive right to use the Licensed Material.

1.2. The rights granted under this Permission Licence are non-transferable and do not include the right to print, sub-license, copy, distribute and/or make available the Licensed Material by any means other than as expressly provided under this Permission Licence.

1.3. Subject to clause 3 below, where electronic files are provided by the Licensor to the Licensee, the rights granted under this Permission Licence do not include the right to store and upload the Licensed Material onto any other Website, Intranet or Database not specifically mentioned in the Licence Cover Sheet.

1.4. The rights granted under this Permission Licence do not include the right to use any third party copyright material incorporated in the Licensed Material. Licensee undertakes to check the Licensed Material carefully and seek permission for use of any such third party copyright material from the relevant copyright owner(s).

1.5. Where the Licensed Material includes videos or audios, the Licensee shall be responsible for payment of all mechanical, synchronisation, communication to the public, public performance and other related fees, royalties and/or other amounts payable to any individual, union, guild, collecting society or similar in the Territory arising in respect of its use of the Licensed Material.

1.6. No changes or adaptations may be made to the Licensed Material without the prior written consent of the Licensor.

1.7. Where the Licensee is permitted to translate the Licensed Material in any language other than English, and as specified in the Licence Cover Sheet, the Licensee shall ensure that any translation of the Licensed Material for the purposes of inclusion in the Licensee's Publication is appropriate for the Territory and is a faithful and accurate translation of the original text. The Licensor accepts no responsibility for any inaccurate translations of the Licensed Material.

1.8. Where the Licensee's publication type is a dissertation, examination paper or photocopy, or where the Licensee is a school, institution or university and is provided with electronic files for the purposes of uploading the Licensed Material on the Licensee's database, interactive white board, website or Intranet, the Licensee's Publication shall not be distributed, stored or sold for any commercial purpose and shall be used solely for the purposes of academic study.

1.9. The Licensed Material shall not appear in the Licensee's Publication in any derogatory context.

2. Copyright Notice and Acknowledgment

2.1. The following acknowledgment shall appear in a sufficiently prominent place on the Licensee's Publication:

Material from: 'AUTHOR, TITLE, published [YEAR] [publisher - as it appears on our copyright page] reproduced with permission of SNCSC'.

2.2. Failure to include the acknowledgment as provided above shall result in termination of this Permission Licence.

3. Reversion of Rights

3.1. The rights granted under this Permission Licence will terminate immediately and automatically upon the earliest of the following events to occur:

3.1.1. Where applicable, the Licence Fee not being received by the Licensor in full by the payment date specified in the relevant invoice;
3.1.2. The Licensed Material not being used by the Licensee within 18 months of the Licence Date;
3.1.3. Expiry of the Term as specified in the Licence Cover Sheet;
3.1.4. The exceeding of the Maximum Print Run or the Maximum Number of Users, as applicable; or

SPRINGER NATURE

PERMISSION LICENCE - GENERAL

LICENCE COVER SHEET

Licence Date: November 5, 2019
Licence Reference: Wenchao Han/Thesis

PARTIES

1. **SPRINGER NATURE CUSTOMER SERVICE CENTER GmbH ("SNCSC")**, acting as a commissionaire agent of Springer Science+Business Media, LLC233 Spring Street, New York, NY 10013, U.S.A. (the "**Licensor**"); and

2. **Wenchao Han** of A3-123A London Regional Cancer Program (800 Commissioners Road East), London, ON N6A 5W9, Canada (the "**Licensee**")

BACKGROUND

This Licence Cover Sheet sets out the principal commercial terms under which the Licensor has agreed to license the Licensed Material (as defined below) to the Licensee for use in the Licensee's Publication. The terms of this Licence Cover Sheet are to be read in conjunction with the General Terms and Conditions, which together with this Licence Cover Sheet constitute the licence agreement (the "**Permission Licence**") between the Licensor and the Licensee as regards the Licensed Material. The terms set out in this Licence Cover Sheet take precedence over any conflicting provision in the General Terms and Conditions. For the avoidance of doubt, derivative rights are excluded from this Permission Licence.

LICENSED MATERIAL

- **Title:** Edge, Stephen B., *AJCC Cancer Staging Handbook* 7th Ed., Springer New York, 2010
- **ISBN:** 978-0-387-88442-4
- **Amount used:** Table *Pathologic*, Chapter "Prostate", p. 461
- **Format Rights:** Electronic rights only
- **Licensee's Publication:** PhD Thesis to be published at Scholar@Western
- **Languages:** English
- **Adaptation rights:** minor editing privileges (stylistic alterations only)
- **Term:** lifetime of webpage of publication (web)
- **Territory:** worldwide
- **Non-exclusive copyright licence:** this is a non-exclusive copyright licence
- **Licence Fee:** gratis

SPRINGER NATURE

3.1.5. Breach by the Licensee of any of the terms included in this Permission Licence.

4. Miscellaneous

4.1. This Permission Licence constitutes the entire agreement between the Licensor and the Licensee and supersedes all communications, negotiations, arrangements and agreements, whether oral or written, between the Licensor and the Licensee with respect to the subject matter of this Permission Licence.

4.2. We do not routinely require a complimentary copy of your product on publication; however, we reserve the right to receive a free copy on request at a later stage.

4.3. If this Permission Licence is for 'reuse in a thesis', electronic rights are granted for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo)

4.4. This Permission Licence shall be governed by, and shall be construed in accordance with, the laws of the Federal Republic of Germany.

4.5. Accessing and using the Licensed Material or any data therein is deemed to be acknowledgment and acceptance in full by the Licensee of the Licence Cover Sheet and these General Terms and Conditions.

Curriculum Vitae

Wenchao Han

Post-secondary Education and Degrees:

- **Ph.D. Candidate in Medical Biophysics** (2014.01–Present). Supervisor: Aaron Ward, Ph.D., Department of Medical Biophysics, The University of Western Ontario, London, Canada
  Prostate cancer detection and grading on mid-gland whole-mount digital histopathology images
- **M.Eng. in Electrical and Computer Engineering** (2012.09–2013.12). The University of Western Ontario, London, Canada
  Communication systems and data networking
- **B.A. Sc. in Optical and Electronics Information** (2006.09–2010.06). Huazhong University of Science and Technology, Wuhan, China

Research Experience

Research Assistant/ PhD Candidate (2013.06–Present)
Lawson Research Institute, Western University, London, Canada

Research Assistant/Undergraduate Student (2010.02–2010.06)
National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China

Research Assistant/Undergraduate Student (2008.03–2009.08)
National Engineering Laboratory for Next Generation Internet Access, Huazhong University of Science and Technology, Wuhan, China

Industry Experience

Project Associate (2010.06–2010.10**)**
FiberHome Telecommunication Technologies Co. Ltd, Wuhan, China

Other Professional Experience

Reviewer (2019)
Journal of Medical Imaging

Poster judge (2019)
London Health Research Day, London, Canada

Educational Talk (2018)
London Regional Cancer Program, London, Canada
What is machine learning/deep learning?

Publications

Journal Papers

1. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A.D. Ward. Automatic cancer detection and localization on mid-gland radical prostatectomy histopathology images. (under review)
2. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A.D. Ward. *Histological tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens*. (under review)
3. W. Han, C. Johnson, M. Downes, T.H. van der Kwast, J. L. Chin, S. E. Pautler, A.D. Ward. *Automatic prostate cancer sub-grading on digital histopathology images of radical prostatectomy specimens.* (in preparation)

Peer Reviewed International Conference Papers

1. W. Han, C. Johnson, M. Downes, T.H. van der Kwast, J. L. Chin, S. E. Pautler, A.D. Ward. *Automatic cancer sub-grading on digital histopathology images of radical prostatectomy specimens.* SPIE Medical Imaging 2020.02 (upcoming oral presentation)
2. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic high-grade cancer detection on prostatectomy histopathology images.* SPIE Medical Imaging, 2019.02. (oral presentation)
3. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic cancer detection and localization on prostatectomy histopathology images.* SPIE Medical Imaging, 2018.02. (oral presentation)

Peer Reviewed International Conference Abstracts

1. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A.D. Ward. *Prostate Cancer Grading on Gigapixel Microscopy Images Using Machine Learning. Journal of Pathology Informatics, upcoming (oral presentation)*
2. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A.D. Ward. *Automatic cancer and high-grade cancer detection and localization on whole-mount digital histopathology images of mid-gland radical prostatectomy specimens.* Pathology Informatics Summit, 2018.05. (oral presentation)
3. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer detection and contouring on digital histopathology imaging.* Pathology Informatics Summit, 2017.05. (oral presentation)

Peer Reviewed Local Conference Abstracts

1. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Finding and grading prostate cancer on gigapixel microscopy images using machine learning.* Oncology Research and Education Day, 2019.06. (poster presentation)
2. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic cancer detection and grading on gigapixel microscopy images of radical prostatectomy tissue sections using machine learning.* London Imaging Discovery Day, 2019.06. (oral presentation)
3. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A.D. Ward. *Automatic high-grade cancer detection on digital histopathology images for prostate cancer.* London Health Research Day, 2019.04. (oral presentation)
4. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic high-grade prostate cancer detection on digital histopathology imaging.* Imaging Network Ontario, 2019.03. (oral presentation, upcoming)
5. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate (high-grade) cancer detection and visualization on digital histopathology imaging.* London Imaging Discovery Day, 2018.06. (oral presentation)
6. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate (high-grade) cancer detection and visualization on digital histopathology imaging.* Oncology Research and Education Day, 2018.06. (oral presentation)
7. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A.D. Ward. *Automatic prostate cancer detection and localization on digital histopathology imaging.* London Health Research Day, 2018.04. (poster presentation)
8. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer detection and localization on digital histopathology imaging.* Imaging Network Ontario, 2018.03. (oral presentation)
9. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer detection and visualization on digital histopathology imaging.* London Imaging Discovery Day, 2017.06. (oral presentation)
10. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer detection and visualization on digital histopathology imaging.* Oncology Research and Education Day, 2017.06. (poster presentation)
11. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer detection and contouring on*

digital histopathology imaging. London Health Research Day, 2017.04. (oral presentation)

12. W. Han, C. Johnson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer detection and contouring on digital histopathology imaging.* Imaging Network Ontario, 2017.03. (oral presentation)
13. W. Han, E. Gibson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer mapping on digital histopathology imaging.* London Health Research Day, 2016.04. (poster presentation)
14. W. Han, E. Gibson, M. Gaed, J. A. Gomez, M. Moussa, J. L. Chin, S. E. Pautler, G. Bauman, A. D. Ward. *Automatic prostate cancer mapping on digital histopathology imaging.* Imaging Network Ontario, 2016.03. (poster presentation)

Presentations

Podium/Oral Presentations

1. **SPIE Medical Imaging 2020 (Upcoming).** *Automatic cancer sub-grading on digital histopathology images of radical prostatectomy specimens.* Houston, US. February, 2020.
2. **Pathology Visions 2019.** *Prostate cancer grading on gigapixel microscopy images using machine learning.* Orlando, US. October, 2019
3. **London Imaging Discovery Day 2019.** *Automatic cancer detection and grading on gigapixel microscopy images of radical prostatectomy tissue sections using machine learning.* London, Canada, June, 2019
4. **London Health Research Day 2019.** *Automatic high-grade cancer detection on digital histopathology images for prostate cancer.* London, Canada. June, 2019.
5. **Imaging Network Ontario 2019**. *Automatic high-grade prostate cancer detection on digital histopathology imaging.* London, Canada. April, 2019.
6. **SPIE Medical Imaging 2019.** *Automatic high-grade cancer detection on prostatectomy histopathology images.* San Diego, US. February, 2019.
7. **Invited talk.** University of Toronto & Sunnybrook Research Institute. *Automatic prostate cancer detection and grading on histopathology using machine learning.* Toronto, Canada. November, 2018.
8. **Talk On Fridays 2018.** Lawson Research Institute. *Automatic cancer detection and grading using machine learning: comparing conventional machine learning to deep learning.* London, Canada, November, 2018.
9. **London Imaging Discovery Day 2018**. *Automatic prostate (high-grade) cancer detection and visualization on digital histopathology imaging using machine learning.* London, Canada, June, 2018.[*] **(Award)**

10. **Cellular and Molecular Imaging Symposium 2018.** *Automatic prostate cancer detection and visualization on digital histopathology imaging using machine learning.* London, Canada, May, 2018.
11. **Oncology Research and Education Day 2018**. *Automatic prostate (high-grade) cancer detection and visualization on digital histopathology imaging using machine learning.* London, Canada, June, 2018.
12. **Pathology Informatics 2018**. *Automatic cancer and high-grade cancer detection and localization on whole-mount digital histopathology images of mid-gland radical prostatectomy specimens.* Pittsburgh, US. May, 2018.
13. **Imaging Network Ontario 2018**. *Automatic cancer detection and localization on digital histopathology images.* Toronto, Canada. April, 2018.
14. **SPIE Medical Imaging 2018.** *Automatic cancer detection and localization on prostatectomy histopathology images.* Houston, US. February, 2018.
15. **Imaging Applications in Prostate Cancer 2017.** *Automatic cancer detection and localization on prostatectomy histopathology images.* London, Canada. November, 2017.
16. **Pathology Informatics 2017.** *Automatic cancer detection and localization on whole-mount digital histopathology images*. Pittsburgh, US. May, 2017.
17. **London Imaging Discovery Day 2017**. *Automatic cancer detection and visualization on digital histopathology images.* London, Canada, June, 2017.
18. **London Health Research Day 2017.** *Automatic prostate cancer detection and contouring on digital histopathology imaging*. London, Canada. March, 2017.
19. **Imaging Network Ontario 2017.** *Automatic prostate cancer detection and contouring on digital histopathology imaging.* London, Canada. March, 2017.

Poster Presentations

1. **Oncology Research and Education Day 2019** *Finding and grading prostate cancer on gigapixel microscopy images using machine learning.* London, Canada. June, 2019.
2. **OICR Annual Meeting 2019** *Automatic high-grade prostate cancer detection on digital histopathology imaging*. Toronto, Canada. April, 2019.
3. **London Health Research Day 2018.** *Automatic prostate cancer detection and localization on digital histopathology images.* London, Canada. May, 2018.
4. **Oncology Research and Education Day 2017.** *Automatic prostate cancer detection and visualization on digital histopathology imaging.* London, Canada. June, 2017.
5. **OICR Annual Meeting 2017.** *Automatic prostate cancer detection and contouring on digital histopathology imaging.* Toronto, Canada. April, 2017.
6. **Imaging Applications in Prostate Cancer 2016.** *Automatic prostate cancer mapping on digital histopathology imaging.* London, Canada. November, 2016.
7. **Cancer Care Ontario Research 2016.** *Automatic prostate cancer mapping on digital histopathology imaging.* Toronto, Canada. April, 2016.

8. **Imaging Network Ontario 2016.** *Automatic prostate cancer mapping on digital histopathology imaging.* Toronto, Canada. March, 2016.

Research Funding and Awards

1. Lawson Travel Award (2018, SPIE Medical Image conference)
2. NSERC Computer-Assisted Medical Intervention Training Program (2014–2015, CAD $12,000 award)
3. Excellent Graduate Award (2010, based on GPA and employment)
4. Second Thesis Award (2010, the best thesis award in the department)

Training and Certificates

Training

Western University, London, Canada
Lawson Health Research Institute, London Health Science Centre

- WHMIS safety training (2014)

Huazhong University of Science and Technology, Wuhan, China
State Key Laboratory of Laser Technology

- Introduction, operation and application of gas laser and solid laser (2010)

Optical Network Laboratory for SDH/ (MSTP)

- Observation, testing and engineering design of SDH (Synchronous Digital Hierarchy) system (2010)
- Configuration of broadband networking (2010)

Simulation Laboratory for Information Systems

- Circuit VHDL, MCU, FPGA, etc. design and simulation (2009)

Laboratory of optoelectronic and optical fiber technologies

- Optical devices test (2009)
- Configuration and operation of optical communication system (2009)

Certificates

- Certified Rescue Diver (Certified by PADI, 2016)
- Certified First Aid and CPR (Certified by Red Cross Association, 2015)
- Certified Operator for Optical System (Certified by National Department of Labour, 2010)

Extracurricular Roles

Certified Rescue Diver (2014-2017)

- Participated more than 50 dives in more than 20 different dive sites and received 4 diving certificates from PADI, Dominica Republic, Mexico, and Costa Rica.

Mixed Martial Arts Trainee (2014-2017)

- Trained by professional UFC athlete, Chris Clements, Adrenaline Training, London, Canada

Founder and Co-chair (2009–2010)

- Department of Life Advisor, Huazhong University of Science and Technology, Wuhan, China

Member (2008–2010)

- Student Union Committee, Department of Literature and Art, Huazhong University of Science and Technology, Wuhan, China