QUANTITATIVE HISTOMORPHOMETRY OF DIGITAL PATHOLOGY AS A COMPANION DIAGNOSTIC: PREDICTING OUTCOME FOR ER+ BREAST CANCERS

BY AJAY NAGESH BASAVANHALLY

A Dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

and

The Graduate School of Biomedical Sciences

University of Medicine and Dentistry of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Biomedical Engineering

written under the direction of

Anant Madabhushi

and approved by

New Brunswick, New Jersey

January, 2014

ABSTRACT OF THE DISSERTATION

Quantitative Histomorphometry of Digital Pathology as a Companion Diagnostic: Predicting Outcome For ER+ Breast Cancers

by Ajay Nagesh Basavanhally

Dissertation Director: Anant Madabhushi

This work involves the creation of an image-based companion diagnostic framework that employs quantitative features extracted from whole-slide, H & E stained digital pathology (DP) images to distinguish patients based on disease outcome, with a clinical application aimed at distinguishing estrogen receptor-positive (ER+) breast cancer (BCa) patients with good and poor outcomes. Quantitative histomorphometry (QH) – the conversion of a digitized histopathology slide into a series of quantitative measurements of tumor morphology – is a rapidly growing field aimed at introducing advanced image analytics into the histopathological workflow. The thrust towards personalized medicine has led to the development of companion diagnostic tools that measure gene expression, yielding quantitative outcome predictions for improved disease stratification and customized therapies, e.g. Oncotype DX (Genomic Health, Inc.) for ER+ BCa. Yet, tumor morphology is often correlated with genomic assays, suggesting that genotypic variations in biologically distinct classes of tumors lead to distinct patterns of tumor cell morphology and tissue architecture in histopathology.

The application of this work to ER+ BCa is highly relevant to current clinical needs. Current treatment guidelines recommend that the majority of women with ER+ BCa

receive chemotherapy in addition to hormonal therapy; yet, approximately half will not benefit from chemotherapy while still enduring its harmful side effects. Hence, there is a clear need for the development of automated prognostic tools to identify women with poorer outcomes who will likely benefit from chemotherapy.

The primary novel contributions of this work are (1) a color standardization system for improving the consistency in appearance of tissue structures across images, (2) the identification of tissue structures and corresponding QH signatures with prognostic value in ER+ BCa, (3) a multi-field-of-view framework for robust integration of prognostic information across whole-slide DP images, and (4) a method for predicting classifier performance for a large data cohort based on the availability of limited training data. This work will pave the way for the development of novel companion diagnostic systems capable of producing quantitative and reproducible image-based risk scores. These risk scores will play a vital role in decision support by helping clinicians predict patient outcome and prescribing appropriate therapies.

Preface

This dissertation represents the collective works of the author over the course of his thesis work. It is primarily composed from the content of published peer-reviewed journal manuscripts [1–3] and peer-reviewed conference papers [4–8]. Additional content includes work from journal manuscripts that are currently under review. Other papers co-published by the author that are not included in this thesis are available on the website for the Center for Computational Imaging and Personalized Diagnostics.

Acknowledgements

The work performed during the course of this dissertation would not have been possible without the help and support of many people. First, and foremost, I would like to thank Dr. Anant Madabhushi for his guidance over the past seven years. His ability to inspire students and foster collaboration across disciplines has made me the researcher I am today. I would also like to thank my committee members for their invaluable advice throughout my dissertation work.

I have benefited greatly from the strong collaborative effort fostered by my colleagues at the former Laboratory for Computational Imaging and Bioinformatics at Rutgers University and the current Center for Computational Imaging and Personalized Diagnostics at Case Western Reserve University. The tightly woven, overlapping nature of our research would assuredly not have been possible without their efforts. Additionally, I have had the good fortune of collaborating with world-class clinicians over the last seven years, including Dr. Shridar Ganesan, Dr. John Tomaszewski, Dr. Michael Feldman, Dr. Hannah Gilmore, Dr. Natalie Shih, and Dr. Carolyn Mies.

The work performed in this dissertation was funded in part by the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (Grant Nos. R01CA136535-01, R01CA14077201, R21CA127186 01, and R03CA143991-01) and The Cancer Institute of New Jersey.

Above all, I would like to thank my parents, Nagesh and Jayanthi, for instilling in me a love of science and a desire to address the problems that affect people's lives. Words cannot express the impact of the support and encouragement offered by my parents and my brother, Naveen, throughout the course of this dissertation.

Dedication

To my parents – for making me the person I am today.

It always seems impossible until it's done.

– Nelson Mandela

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Quantitative histomorphometry and digital pathology

The rise of quantitative histomorphometrics (QH) – the conversion of a histopathology slide into a series of quantitative measurements of tumor morphology – is closely coupled with the advent of digital pathology (DP), a rapidly growing field that includes the computerized scanning, visualization, and analysis of tissue specimens generated by various surgical procedures (e.g. biopsy, resection). Early applications of quantitative DP involved basic measurements such as cell counting, object size measurement, and light absorption characteristics [9]; yet the development of complex image analytics was limited by the lack of high-resolution sensors, storage space, and computational throughput. Recent technological advances have led high-throughput, high-resolution, whole-slide DP scanners to become increasingly commonplace in the clinical setting for both primary diagnosis and telepathology [10]. Quantitative analysis of DP specimens is starting to gain popularity in clinical practice, but is often limited to identifying staining extent in IHC-stained DP images. Routine hematoxylin and eosin (H & E) staining, which enhances visualization of tissue differentiation and tumor morphology, accounts for the vast majority of histopathology studies. However, the development of robust QH tools for large-scale diagnostic and prognostic analysis of H & E stained DP has proven to be difficult [11].

1.2 Role of QH in personalized medicine

Recent advances in personalized medicine – the idea that disease diagnosis and treatment can be tailored to fit the individual needs of each and every patient – are rapidly changing many aspects of patient care today. In particular, the development of molecular assays that measure gene and protein expression, referred to collectively as *companion diagnostics*, have led to quantitative outcome predictions that improve disease stratification and allow for more customized therapies. These assays, however, suffer from a number of translational drawbacks including high cost, the need for specialized facilities, and increased time to treatment. Furthermore, recent studies have demonstrated correlations between molecular assays and tumor morphology descriptors, suggesting that the linkage between molecular assays and patient outcome is not unique [12–14]. In this work, we hypothesize that the genotypic changes measured by gene expression assays are also reflected by variations in tissue morphology, which can be characterized by QH analysis in a quantitative and reproducible manner. Thus, a unified framework able to integrate different types of quantitative image features across all aspects of whole-slide DP images will play an important role in the development of QH-based companion diagnostic systems.

1.3 Application of QH to breast cancer outcome prediction

The majority of the work presented in this thesis is applied to breast cancer (BCa), which is the most common cancer diagnosis for women in the United States with an annual incidence of invasive malignancies greater than 200,000 and mortality greater than 40,000 [15]. Measurement of estrogen receptor (ER) expression is a routine part of the clinical evaluation due to the availability of targeted therapies (e.g. tamoxifen). Current treatment guidelines recommend that the vast majority of women with tumors expressing the estrogen receptor (i.e. ER+ BCa) receive chemotherapy in addition to the normally prescribed hormonal therapy; yet, the majority will not benefit from chemotherapy while enduring the harmful side effects. Hence, there is a clear need for the development of automated prognostic tools to identify women with poorer outcomes (i.e. more aggressive cancers) who are likely to derive the greatest benefit from chemotherapy.

In current clinical practice, the Oncotype DX gene expression assay (Genomic Health, Inc.) is commonly used to yield a quantitative Recurrence Score (RS) for lymph node-negative (LN-), ER+ BCa patients that has shown to be correlated with disease outcome and response to therapy [16]. In fact, an analogous relationship between QH and disease outcome can be inferred by acknowledging the similarities between QH analysis and BCa tumor grade (e.g. characterization of tumor differentiation and nuclear pleomorphism). Studies have previously demonstrated that both high tumor grade and high Oncotype DX RS are significant predictors of tumor recurrence; yet, the two predictors are not independent of one another [17]. Indeed, genes that tend to drive a high RS, such as those associated with cell proliferation and HER2 amplification, are also associated with high grade tumors. Further studies have also confirmed the high degree of correlation between tumor grade and RS [12–14], suggesting that the visual information present in standard H & E histology contains prognostic information similar to that present in gene expression data. Our overall hypothesis is that gene expression signatures from biologically distinct classes of tumors lead to distinct patterns of tumor cell morphology and tissue architecture, and that these patterns can be identified by the computer-aided QH analysis of histological images.

1.4 Surrogate ground truth for patient outcome in ER+ breast cancer

While the ideal ground truth for evaluation of prognostic tools such as the one described in this work is long-term patient outcome (i.e. recurrence-free survival), this type of data is very difficult to obtain. In lieu of patient outcome, we employ both modified Bloom-Richardson (mBR) grading [18] and Oncotype DX Recurrence Score (RS) as *surrogate* ground truth markers for LN-, ER+ BCa patients.

A number of prognostic criteria have been developed to characterize the disease aggressiveness in invasive BCa tumors via visual analysis of H & E stained histopathology, including the Bloom-Richardson [19] and Nottingham grading schemes [18]. In particular, the Nottingham, or modified Bloom-Richardson (mBR), system has gained popularity due to the integration of descriptors that characterize large-scale tissue structure with those that describe nuclear differentiation [20]. The mBR grading system encompasses three visual signatures: degree of tubular formation, nuclear pleomorphism, and mitotic activity. Each signature is scored semi-quantitatively on a scale of 1-3 to produce a combined mBR grade on a scale of 3-9 [18]. For prognostic purposes, patients are commonly split into three classes corresponding to low (mBR 3-5), intermediate (mBR 6-7), and high (mBR 8-9) grades. The Oncotype DX gene expression assay has been clinically validated to predict the likelihood of 10-year distant recurrence and the expected benefit from adjuvant chemotherapy for LN-, ER+ BCa patients [16]. The assay, which yields a quantitative Recurrence Score (RS) between 0 and 100, has been shown to have the predictive power to distinguish low (RS $<$ 18), intermediate (18 $\leq$ RS $\leq$ 30), and high (RS $>$ 30) risk patients.

It is important to note that the close relationship between mBR grade and prognosis (i.e. prediction of patient outcome) is well-known [12, 13]; yet clinical usage of mBR grade is often tempered by concerns about intra- and inter-rater variability [21–23]. Meyer et al. [23] showed that agreement between seven pathologists is only moderately reproducible ($\kappa = 0.50 - 0.59$), while Dalton et al. [22] further illustrated the suboptimal treatment that can result from incorrect mBR grading. Boiesen et al. [21] demonstrated similar levels of reproducibility ($\kappa = 0.50 - 0.54$) across a number of pathology departments. A possible reason for this discrepancy is that pathologists currently lack the automated image analysis tools to accurately, efficiently and reproducibly quantify mBR grade in histopathology.

1.5 Challenges in the development of QH methods for DP analysis

While computerized image analysis tools for DP have become increasingly sophisticated [11], existing frameworks for automated analysis of whole-slide DP are particularly lacking. Current approaches face a number of challenges including (1) intrinsic biological heterogeneity, (2) variable slide preparation and digitization, and (3) computational limitations of large whole-slide DP images that are frequently 10^10 pixels in size.

Breast cancer is known to contain intratumoral heterogeneity on both the genomic [24–26] and morphologic [27, 28] levels, which are highlighted specifically by

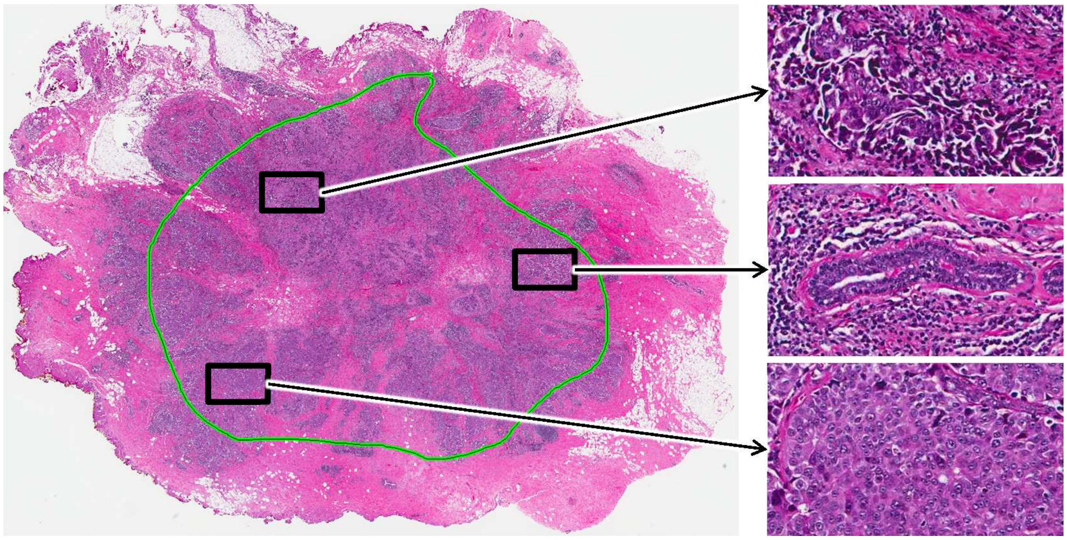


Figure 1.1: FOVs taken from a single histopathology slide illustrate the high level of intratumoral heterogeneity in ER+ BCa. The green annotation represents invasive cancer as determined by an expert pathologist. Note the disorganized tissue structure of some FOVs (top, bottom) represent more aggressive cancers than others (middle).

Gerlinger et al. as a major hurdle in the development of personalized therapeutic strategies [26]. In histopathology imagery, this challenge is exemplified routinely by the coexistence of regions containing cancerous and non-cancerous tissue, different types of cancer (e.g. ductal carcinoma *in situ* and invasive ductal cancer), and levels of tumor differentiation (e.g. low and intermediate grades) on a single slide (Figure 1.1). This phenomenon suggests that predictions based on just a few isolated fields-of-view (FOVs) may not accurately reflect the level of disease aggressiveness in an entire DP slide; instead, a more detailed analysis comprising a multitude of FOVs may be needed to understand the true nature of the tumor.

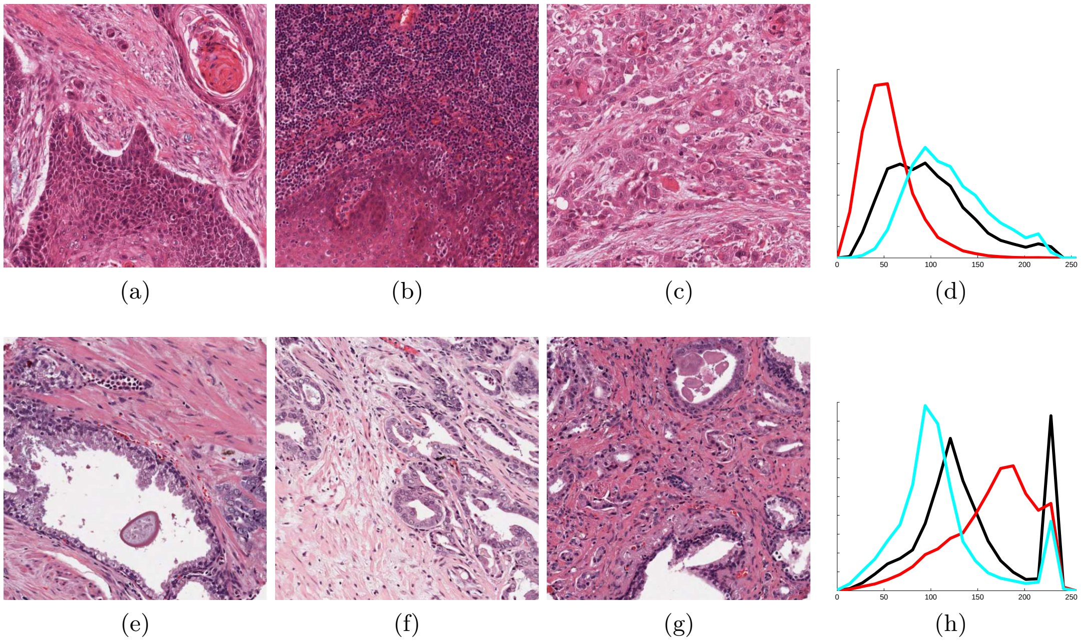
Apart from biological sources of heterogeneity, additional variability is introduced during the slide preparation and digitization processes. Lack of standardization in slide preparation leads issues such as (1) variable staining based on stain manufacturer, batch, and fixation time and (2) tissue thickness based on block preparation and tissue folding. Additional variations in the whole-slide digitization process can be caused by differences in scanning device manufacturers, illumination, compression, and post-processing algorithms.

1.6 QH-based companion diagnostic framework for whole-slide DP analysis

The goal of this thesis is to develop an QH-based companion diagnostic framework for the prediction of disease outcome in early stage, ER+ BCa patients using only QH features extracted from whole-slide DP images.

1.6.1 Color standardization

The development of tools for the processing of color images is often complicated by non-standardness – the notion that different image regions corresponding to the same tissue will occupy different ranges in the color spectrum (Figure 1.2). In DP, these issues are often caused by variations in slide thickness, staining, scanning parameters, and illumination. Nonstandardness can be addressed via standardization, a pre-processing step that aims to improve color constancy by realigning color distributions of images to match that of a pre-defined template image. Unlike color normalization methods, which aim to scale (usually linearly or assuming that the transfer function of the system is known) the intensity of individual images, standardization is employed to align distributions in broad tissue classes (e.g. epithelium, stroma) across different DP images irrespective of institution, protocol, or scanner. Intensity standardization has previously been used for addressing the issue of intensity drift in MRI images, where similar tissue regions have different image intensities across scanners and patients. By contrast, histopathological imagery is complicated by the (a) additional information present in color images and (b) heterogeneity of tissue composition. In this work, we present a novel Expectation Maximization-based segmentation-driven color standardization (EMS) scheme to decompose histological images into independent tissue classes (e.g. nuclei, epithelium, stroma, lumen) via the EM algorithm and align the color distributions for each class independently. In addition to the flexibility offered by this approach, EMS is more suited for the analysis of retrospective data because it does not require prior information about the staining and digitization processes.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1.2: H & E stained histopathology images of (a)-(c) oropharyngeal and (e)-(g) prostate cancers demonstrate color nonstandardness across tissue specimens resulting from variations in slide preparation (e.g. tissue thickness and staining). These variations are reflected by (d), (h) corresponding histograms of the green color channel, in which each image occupies different ranges of the color spectrum.

1.6.2 Detection of relevant tissue structures

The identification of various tissue structures provides the basic building blocks for quantifying tissue architecture and tumor morphology. While the majority of detection and segmentation tasks have focused on low-level objects (e.g. nuclei, stroma, lumen), the detection of more complex tissue structures may play an important role in extracting prognostic QH features. An important criterion for identifying complex objects with multiple attributes is the use of domain knowledge which reflects the precise spatial linking of the constituent attributes. Hence, simply detecting the presence of the low-level attributes that constitute the object, even in cases where these attributes might have spatial proximity to each other, may not be a robust strategy. The O'Callaghan neighborhood [29] is an ideal vehicle for characterizing objects comprised of multiple attributes spatially connected to each other in a precise fashion because it allows for modeling and imposing spatial distance and directional constraints on the object attributes. In this work we apply the O'Callaghan neighborhood to the problem of tubule identification on H & E stained BCa histopathology, where a tubule is characterized by a central lumen surrounded by cytoplasm and a ring of nuclei around the cytoplasm. The detection of tubules is important in ER+ BCa because tubule formation is an important component of the mBR grading system, which is strongly linked to disease aggressiveness and patient outcome. The more standard pattern recognition approaches to detection of complex objects typically involve training classifiers for low-level attributes individually. For tubule detection, the spatial proximity of lumen, cytoplasm, and nuclei might suggest the presence of a tubule. However such an approach could also suffer from false positive errors due to the presence of fat, stroma, and other lumen-like areas that could be mistaken for tubules. Instead, we identify tubules by taking advantage of the distance and spatial constraints imposed by the construction of O'Callaghan neighborhoods comprised of nuclei around each luminal area.

1.6.3 Extraction of prognostically relevant QH features

We explore multiple classes of QH features that characterize the (1) 2D spatial arrangement of multiple nuclei, (2) textural variations within individual nuclei, and (3) degree of tubule formation in DP images, all of which reflect various aspects of the mBR grading system. The spatial arrangement of nuclei (i.e. nuclear architecture) is used to model the overall level of tissue disorder in an image. In this work, we quantify this concept by using individual nuclei as vertices for the construction of various graphs (Voronoi graph, Delaunay triangulation, and minimum spanning tree) and, subsequently, extract relevant statistics related to the size, shape, and length of these graphs [30]. Another way to quantify tissue disorder is by quantifying tubule formation, a key component of the mBR grading system [18]. Hence, in this work we utilize our ability to identify tubules and define QH features that quantify the degree of tubule formation in BCa DP images (Figure 1.3). Textural patterns within nuclei (i.e. nuclear texture) are used to model the intra-nuclear variations in chromatin arrangement, which is generally more heterogeneous in rapidly dividing, higher grade nuclei [31]. In this work, we employ second-order Haralick statistics are calculated from gray-level co-occurrence matrices

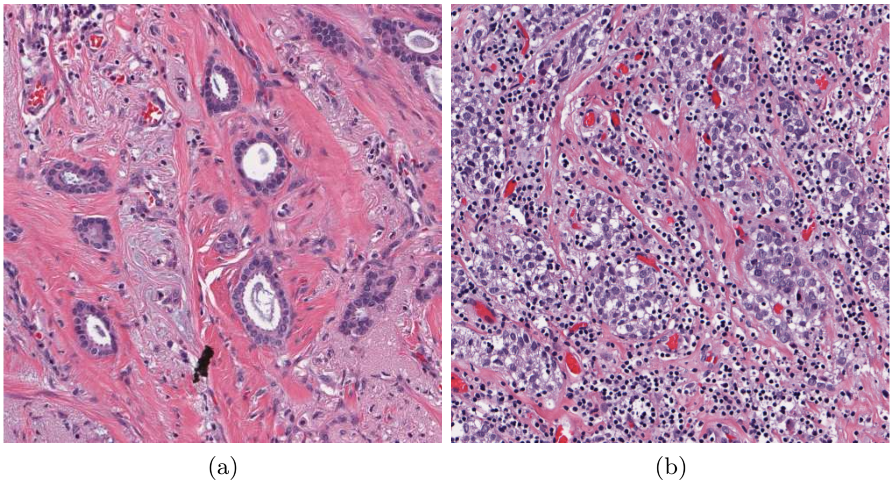


Figure 1.3: Breast cancer histopathology images corresponding to (a) low and (b) high tubule subscore, a key component of mBR grading. Tissue with a low tubule subscore has a higher proportion of nuclei arranged in tubules and corresponds to better outcomes.

within segmented nuclear regions [32].

Conceptually, a large number of descriptive features is highly desirable in terms of distinguishing patients based on mBR grade. In reality, however, large feature sets present problems in data classification such as (1) the curse of dimensionality [33], which calls for an exponential growth in the data cohort for each additional feature used, (2) the inability to identify specific features containing class discriminatory information, and (3) the presence of redundant features that do not provide any additional information to the classifier. In this work, we mitigate these challenges by employing the Minimum Redundancy Maximum Relevance (mRMR) feature selection scheme [34]. Given a a set of samples with corresponding features and class labels, the mRMR algorithm identifies the most relevant features by simultaneously maximizing mutual information between the features and class labels (i.e. maximizing relevance) and (2) minimizing mutual information between individual features (i.e. minimizing redundancy).

1.6.4 Whole-slide classification via the multi-FOV framework

QH features extracted in this work are used in conjunction with a novel multi-field-of-view (multi-FOV) classifier – a whole-slide classifier that extracts features from a multitude of FOVs of varying sizes – to distinguish ER+ BCa patients based on predicted disease outcome [2, 3, 7]. The multi-FOV scheme uses a fixed image resolution and extracts image features at FOVs of different sizes, a highly desirable attribute for extracting QH descriptors from heterogeneous images where it is not clear which FOV sizes will contain class discriminatory information. First, a slide is split into FOVs of a fixed size and relevant image features are extracted. A pre-trained classifier makes an initial class decision for each FOV and the decisions for all FOVs are aggregated to make a single class prediction for the specific FOV size. This procedure is repeated for a variety of FOV sizes and the class predictions at all FOV sizes are aggregated to arrive at a single decision for the entire slide. Hence there is no need to empirically determine the optimal FOV size for classification; rather this approach extracts QH descriptors across many FOV sizes in parallel and combines their class predictions to form a meta-classifier.

1.6.5 Selecting an appropriate classifier based on limited training data

Although the multi-FOV approach employs a pre-trained classifier, the selection of an optimal classifier given only a small dataset is not straightforward. Clinical trials increasingly employ medical imaging data in conjunction with supervised classifiers, where the latter require large amounts of training data to accurately model the data. [35–37]. Yet, a classifier is often selected at the start of the trial based on smaller and more accessible datasets that are not sufficiently generalizable, thus yielding inaccurate and unstable classification performance [37, 38]. We aim to address two common concerns in classifier selection for clinical trials: (1) predicting expected classifier performance for large datasets based on error rates calculated from smaller datasets and (2) the selection of an appropriate classifier based on expected performance for large datasets that will be available in the future [39]. The selection of an optimal classifier for a specific dataset usually requires large amounts of annotated training data [40] since the error rate of a supervised classifier tends to decrease as training set size increases [41]. However, in clinical trials, this decision is often based on the assumption (which may not necessarily hold true [5]) that the relative performance of classifiers on a smaller dataset will remain the same as more data becomes available.

In this work, we aim to overcome the major constraints on classifier selection in clinical trials that employ medical imaging data, namely (1) the selection of an optimal classifier using only a small subset of the full cohort and (2) the prediction of long-term performance in a clinical trial as data becomes available sequentially over time. We present a framework for comparative evaluation of classifiers using only limited amounts training data by using random repeated sampling (RRS) in conjunction with a cross-validation sampling strategy. First, the dataset is split into K distinct pools where one pool is used for testing while the remaining $K-1$ are used for training. A subsampling procedure is used to create multiple subsets of various sizes from the training pool. Each subset is used to train a classifier, which is then evaluated against the testing pool. The pools are rotated K times to ensure that all samples are evaluated once, after which all error rates are averaged for each training set size. The resulting mean error rates are used to determine the three parameters of the power-law model (rate of learning, decay rate, and Bayes error) that characterize the behavior of error rate as a function of training set size.

1.7 Primary Goals of this thesis

In summary, the work described in this thesis comprises 4 goals that encompass the range of tasks needed to develop a QH-based companion diagnostic framework.

1. Standardization of DP images due to variations in slide preparation and scanning hardware.

2. Detection of higher-order tissue structures in DP images with a specific focus on identifying tubule formation in BCa histopathology.

3. Development of theoretical and biological intuition for a multi-FOV framework that integrates sampling, feature extraction, and classification of whole-slide DP.

4. Prediction of large-scale classifier performance and optimal classifier selection based on the availability of limited training data.

1.8 Organization of this thesis

The organization of this thesis is as follows. Chapter 2 reviews the existing literature and details the novel contributions for each goal. Chapters 3-7 describe the methods developed and experiments used to achieve each of the primary goals of this thesis: (a) standardization of DP images, (b) detection and of various tissue structures and subsequent feature extraction, (c) development of the multi-FOV framework for whole-slide DP analysis, and (4) prediction of large-scale classifier performance from smaller data cohorts. In Chapter 8, we present the experimental design followed by results and discussion in Chapter 9. Chapter 10 summarizes the main achievements of this work with concluding remarks and suggested directions for future research are presented in Chapter 11.

Chapter 2

Previous work and novel contributions

Although the concept of an entirely image-based decision support framework for whole-slide DP is a recent development, there exists a body of relevant work for the individual components of the framework. In this section, we discuss both the previous approaches to each of the primary goals laid out in Section 1 as well as the major novel contributions of this work.

2.1 Approaches to standardization in medical image analysis

The development of standardization techniques for biomedical imaging data is driven by need to maintain intensity or color constancy across multiple images in a cohort. For instance, computerized analysis of MR images is often complicated by intensity drift, where multiple images acquired from the same scanner occupy different ranges in the intensity spectrum [42–44]. Methods to correct intensity drift include a piecewise intensity standardization approach that employed linear interpolation to define landmarks at evenly spaced percentiles of an intensity distribution [42]. The distribution for a new *test* image was standardized to a pre-defined *template* image by shifting the intensity distribution of the test image to match that of the template image between each corresponding set of landmarks [42]. Madabhushi et al. [43] further extended the piecewise standardization approach by implicitly incorporating basic spatial information via the generalized scale model. This method, however, does not easily translate to DP images for a number of reasons. First, their approach was limited to a connected component labeling that (1) has no particular correspondence between DP images and (2) cannot be used for tissue classes (e.g. nuclei) spread across many regions. In addition, intensity standardization employs global distribution alignment using a single histogram to characterize an entire image [42, 43]. This type of global standardization (GS) is unable to account for the heterogeneity of DP images, which contain broad, independent tissue classes (e.g. stroma, epithelium, nuclei, lumen) in varying proportions, leading to skewed color distributions and errors in the standardization process [8].

Previous work in maintaining color constancy in DP images has traditionally employed normalization and calibration techniques. Normalization, the process by which color distributions of images are adjusted to fit a predetermined range, is performed independently for each image using only image information available in the image itself [1, 45, 46]. For example, Ballerini et al. [46] normalized images containing non-melanoma skin lesions by using regions of normal skin found within the same image. Others have taken a more implicit approach to normalization by operating in alternate color spaces (e.g. HSV, CIE-Lab) that are more invariant to the effects of color variations [1, 45]. Limitations of normalization include (1) the need for the presence of a "normal" variant within an image and (2) difficulties in accounting for the non-linear intensity variations arising from the many sources of color nonstandardness. Color calibration refers to the modification of acquisition or visualization settings based on prior knowledge of imaging parameters. For instance, Yagi et al. [47] performed calibration of computer monitors for optimal viewing of DP; yet calibration is unfeasible for the analysis of retrospective studies on existing data cohorts. Note that the larger body of work aimed at correcting variations in illumination for images formed by reflective light (e.g. digital photography) is inappropriate for DP images that are formed instead by light absorption [48, 49].

2.2 Detection and segmentation of tissue structures in H & E stained DP

Detection and segmentation of tissue structures in DP images are fundamental to the subsequent extraction of quantitative, reproducible, and clinically relevant image features. Due to the importance of characterizing tissue architecture for the diagnosis and treatment of various cancers, the majority of previous work in object detection has focused on the delineation of low-level structures (e.g. nuclei, stroma, lumen) [1,4,50–53]. Prior to the construction of an O'Callaghan neighborhood, we must first identify (1) all cancer nuclei and (2) all potential lumen areas within the image. Previous approaches to automated nuclear detection have also relied on differences in staining to distinguish nuclei from surrounding tissue, including fuzzy c-means clustering [51], adaptive thresholding [50], Expectation-Maximization [4], and region-growing [1] methods. However, these methods are often highly sensitive to initial values and parameter selection.

The segmentation of white luminal areas is a key component of identifying glands in DP images. Methods such as Bayesian classifiers [54] and fuzzy clustering [55] have previously been used for lumen segmentation, but may not be appropriate since they often require large amounts of training data or exhibit high sensitivity to initialization. Other techniques such as region growing [52] have successfully been used to identify lumen in prostate cancer histopathology; however, they require image intensity within the lumen areas to be homogeneous and these methods have difficulty handling scenarios where tissue may be interspersed within the lumen (Figure 2.1(a)). Traditional boundary-based active contour models are often limited by high sensitivity to initial positions [53, 56]. However, Xu et al. [56] employed a Hierarchical Normalized Cut (HNCut) initialized Color Gradient based Active Contour (CGAC) model to segment luminal areas, producing improved performance over traditional active contour methods by incorporating a more robust initialization via HNCut [57] and a more informative color gradient model.

In comparison to low-level tissue structures, there has been limited work towards the identification and characterization of complex, multi-attribute objects, e.g. tubules and glands, comprised of two or more low-level structures. Hafiane et al. [55] used a combination of fuzzy c-means clustering with spatial constraints to identify and segment glandular structures in prostate cancer histopathology, but these techniques are often too sensitive to the presence of outliers. Naik et al. [54] also segmented prostate glands by integrating pixel-level, object-level, and domain-specific relationships via Bayesian classifiers. Probabilistic methods, however, require large amounts of training data to

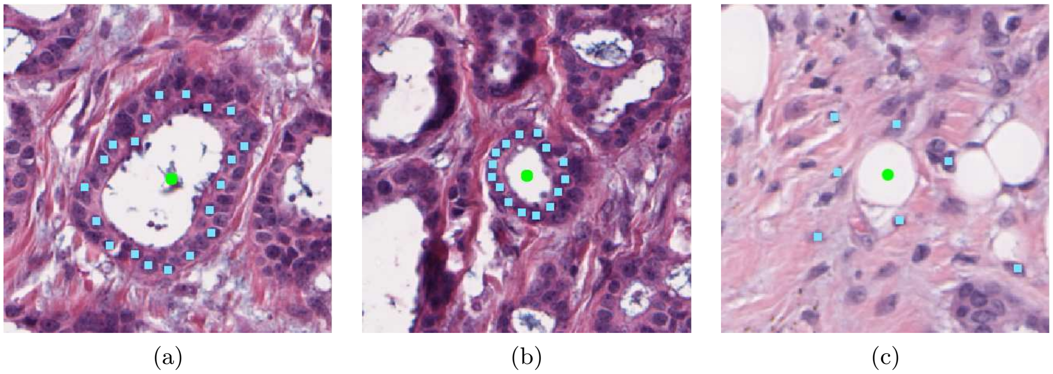


Figure 2.1: Three potential lumen corresponding to (a) a large tubule, (b) a small tubule, and (c) adipose tissue are represented by their respective centroids (green circle) and O'Callaghan neighborhoods (blue squares). By calculating statistics describing the arrangement of the O'Callaghan neighborhoods, (a) and (b) can successfully be classified as tubules while (c) will be correctly identified as a non-tubule object.

accurately model the prior distribution and perform poorly when new data does not fit the trained model. Previously, Kayser et al. [58] have shown the effectiveness of using O'Callaghan neighborhoods to understand the spatial relationships between glands in colon mucosa. By treating individual glands as vertices and modeling the connections between glands as edges, a variety of graph-related features were found to separate tissue classes.

2.3 Extraction of relevant QH features from DP images

In most cancers, the relationship between the visual appearance of tumor morphology and patient outcome is governed by a grading system that characterizes the differentiation of a tissue specimen in a semi-quantitative manner. Hence, it is not surprising that computerized feature extraction approaches have focused on reproducing the specific attributes that comprise these grading systems.

Variations in nuclear architecture (i.e. the 2D spatial arrangement of cancer nuclei in histopathology) are important in clinical practice because they allow pathologists to distinguish between normal and cancerous tissues as well as different levels of tumor differentiation. In the mBR grading system for breast cancers, this is exemplified by characterization of tubule formation in histopathology [18]. A popular approach to modeling nuclear architecture is via the construction of graphs in which individual tissue structures (e.g. nuclei) are defined as vertices and connected by edges. Relevant statistics related to the size, shape, and length of these graphs are then extracted to quantify the image. Such graph-based features have previously been used to accurately distinguish variations in lymphocytic infiltration [1], cancer type [59], tumor grade [4, 30], and prognosis [4] in digitized BCa histopathology, as well as hierarchical tissue structure in glioma [60] and tumor grade in prostate [61]. In addition, researchers have recently demonstrated the ability to identify high grade regions within individual BCa histopathology slides via sparse analysis of Voronoi graphs [62]. More recently, Ali et al. used small clusters of nuclei as vertices to construct cell cluster graphs for predicting biochemical recurrence from prostate cancer DP images [63]. The applicability of graph-based features for a wide range of diseases and classification tasks suggests that they are able to quantify the general large-scale patterns that reflect varying levels of tissue organization across different disease states.

Textural information from nuclear regions (i.e. nuclear texture) represents the variation in chromatin arrangement [31], which is generally more heterogeneous in rapidly-dividing, higher grade BCa cells. The diagnostic importance of nuclear texture in histopathology has been widely studied [51, 64–66]; yet recent work in differentiating BCa grade via analysis of nuclear texture has been limited. For example, Weyn et al. performed a limited study that explored the ability of wavelet, Haralick, and densitometric features to distinguish nuclei from low, intermediate, and high BCa tissues [31]. More recently, Petushi et al. [51]. found that the extent of cell nuclei with dispersed chromatin is related to BCa tumor grade. Note that this differs from studies that have relied on the extraction of textural statistics from entire FOVs (i.e. global texture) [30, 45]. Doyle et al. utilized grey-level, Gabor, and Haralick texture features extracted from entire FOVs to discriminate low and high grade tumors in both prostate [45] and BCa [30] histopathology. In addition, Haralick features [32] (i.e. second-order statistics calculated from a gray-level co-occurrence matrix) have previously been used in both nuclear and global textural analysis for classification of tumor grade in numerous cancers, including the breast [31,51], prostate [45], and thyroid [66].

Recently, researchers have also explored the use of fractals to describe the variations architectural complexity of epithelial tissue with respect to the level of differentiation of cells in BCa tumors [67–70]. While these studies are extremely promising, their results are still preliminary because evaluation has generally been limited to isolated FOVs (e.g. individual cells in [68] and TMAs in [69]), relatively small cohorts [68], and specialized stains [69].

2.4 Feature selection and dimensionality reduction approaches

Existing methods developed to reduce the size of large features sets can be split into two broad categories: dimensionality reduction (DR) and feature selection. Linear DR methods such as principal component analysis and multidimensional scaling perform a linear mapping from the original feature space to a lower dimensional space, i.e. each dimension in the reduced space is defined by some linear combination of the original features. Previous research illustrating the intrinsic non-linearity of biomedical data [71] suggests that non-linear DR methods such as locally linear embedding, Graph Embedding, or Isomaps) may be more appropriate. However, due to the non-linear mapping from the original feature space, it is extremely difficult to identify the specific features used to create the reduced feature space. For the analysis of QH features in BCa DP imagery, DR techniques such as Graph Embedding have previously been used for in applications of cancer detection [30], cancer grading [30], and characterization of lymphocytic infiltration in BCa [1]. By contrast, feature selection methods do not identify the best combination of features; rather, they are supervised approaches that will rank individual features based on their ability to distinguish class labels. Techniques such as Adaboost [72] and Minimum Redundancy Maximum Relevance [34] are frequently used for the reduction of large feature sets. Feature selection approaches have also been utilized to identify of salient features in biomedical data, including the use of Adaboost for prostate cancer detection in DP [45] and Minimum Redundancy Maximum Relevance for prostate cancer detection in MRI [73], salient genes in microarray data [74], and insight into drug interactions of various protein groups [75].

2.5 Sampling approaches for whole-slide DP images

Another challenge that must be addressed in the analysis of whole-slide DP is the overall size of the data, where glass slides are routinely digitized at high spatial resolution (up to 0.25 µm/pixel) and produce very large images containing 10^{10} pixels. Previous work in computerized DP analysis has traditionally avoided these issues via empirical selection of individual FOVs at a fixed size [1, 4, 11, 30, 51, 76]. While this approach is useful for specific research applications, user intervention in a clinical setting may lead to (1) poor reproducibility due to the bias introduced by varying levels of expertise and (2) increased cost in terms of both time to diagnosis and money. Note that manual FOV selection is also intrinsic to tissue microarray (TMA) analysis, in which small tissue "spots" are sampled from larger regions of interest by an expert pathologist [77].

Some researchers have employed random sampling in an effort to address the bias associated with manual FOV selection [31,68,78]. However, classification results based on randomized FOV selection may still suffer from poor reproducibility, especially in heterogeneous cancers [24,26] such as BCa where individual FOVs may not be representative of the overall tumor. More recently, Huang et al. have approached FOV selection from a holistic perspective through the use of dynamic sampling, which incorporates domain information into the identification of salient regions of interest [62].

Alternatively, hierarchical multi-scale (i.e. multi-resolution) classifiers have also been used for the evaluation of large images [45, 79, 80]. These approaches initially operate at a low spatial resolution before proceeding to incrementally higher resolutions. The hierarchical approach increases computational efficiency by ensuring that only relevant data is exposed to the classifier at higher resolutions via predictions made at the previous level (i.e. at a lower resolution). While hierarchical analysis has previously been applied to neuroblastoma [79] and prostate cancer [45] histopathology, limitations include a serialized processing pipeline in which evaluation of higher resolutions is dependent on results first calculated from lower resolutions. In addition, multi-scale frameworks may have more difficulty in analyzing domain-specific image architecture, e.g. QH features describing the spatial arrangement of nuclei, since it remains invariant to changes in scale (although our visual perception and ability to detect objects within the image will vary).

2.6 Classification and stratification methods for DP analysis

Computer decision support systems for histopathology have utilized a variety of approaches for classification and stratification of DP images. Commonly used classifiers include Support Vector Machines [30] and Bayesian approaches [45]. Similarly, stratification of QH features has relied on dimensionality reduction techniques and posterior probabilities produced by specific classifiers. For instance, Graph Embedding has previously been used in BCa DP for visualization of QH features stratifying lymphocytic infiltration [1], begin vs. malignant tissue [30], and tumor grade [30]. Evaluation on most of these systems has been limited to small cohorts, and procedures for classifier selection have been either ad hoc or based on comparison studies involving a small number of training samples.

2.7 Prediction of error rates for large data cohorts and selection of an appropriate classifier

The ability to predict the performance of a classifier as dataset size increases is crucial to the development of decision support systems that employ DP images. Traditional power calculations aim to determine confidence in an error estimate using repeated classification experiments [81], but do not readily address the question of how error rate changes as more data becomes available. Also, they may not be ideal for analyzing biomedical data because they assume an underlying Gaussian distribution and independence between variables [82]. Repeated random sampling (RRS) approaches, which characterize trends in classification performance via repeated classification using training sets of varying sizes, have thus become increasingly popular, especially for extrapolating error rates in genomic datasets [39, 82, 83]. Drawbacks of RRS include (1) no guarantee that all samples will be selected at least once for testing and (2) a large number of repetitions required to account for the variability associated random sampling. In particular, traditional RRS may suffer in the presence of highly heterogeneous datasets (e.g. biomedical imaging data [84]) due to the use of fixed training and testing pools. More recently, methods such as repeated independent design and test (RIDT) [85] have aimed to improve RRS by simultaneously modeling the effects of different testing set sizes in addition to different training set sizes. This approach, however, requires allocation of larger testing sets than RRS, thereby reducing the number of samples available in the training set for extrapolation. It is important to note that the concept of predicting error rates for large datasets should not confused with semi-supervised learning techniques, e.g. active learning (AL) [86], that aim to maximize classifier performance while mitigating the costs of compiling large annotated training datasets [38]. Since AL methods are designed to optimize classification accuracy during the acquisition of new data, they are not appropriate for *a priori* prediction of classifier performance using only a small dataset.

2.8 Novel contributions of this thesis

Apart from the overall innovation of an image-based companion diagnostic framework for whole-slide, H & E stained DP images, each of the individual goals presented in this thesis also contains novel contributions to the state-of-the-art.

2.8.1 Maintaining color constancy in tissue structures across DP images

The first goal of this work is to improve color constancy across a population DP images so that specific tissue structures (e.g. nuclei, epithelium, stroma) have a consistent appearance across all images. While normalization and calibration approaches are unsuitable for this task, we hypothesize that standardization techniques are more appropriate. However, the additional challenges posed by color standardization over the intensity standardization used for radiological images is not trivial. For instance, multiple DP images for a single organ/disease type primarily contain the same tissue structures; yet, differences in the proportions of these tissue structures will skew their corresponding color distributions. We will mitigate this issue by using the Expectation-Maximization (EM) algorithm [87] – an unsupervised clustering algorithm that separates image pixels into a pre-determined number of Gaussian components – to partition and standardize different tissue classes independently. Corresponding tissue classes across different images are automatically matched by minimizing pairwise distances between all classes in a specified color space. Subsequently, we believe that piecewise intensity standardization algorithms can be applied independently for each color channel by constructing histograms using pixels from each tissue class of a *test* image and aligning it to the corresponding tissue class in the *template* image.

Hence, the major contributions of the Expectation-Maximization segmentation-driven color standardization scheme (EMS) are that it

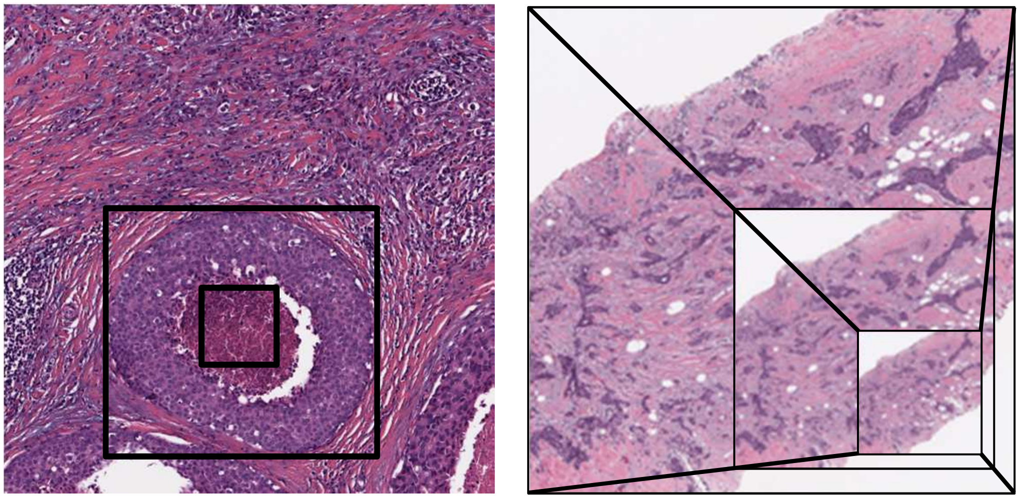
- Aligns color distributions of broad tissue classes (e.g. nuclei, stroma) that are first partitioned via EM; by contrast, previous global methods perform standardization using a histogram of the entire image.
- Can be used retrospectively since it is independent of scanners, staining protocols, and institutions.

2.8.2 Identifying tubules in breast cancer histopathology

The second goal of this work focuses on the detection of higher-order tissue structures that comprise a combination of smaller, low-level attributes. The typical pattern recognition approach for detecting a multi-attribute object O would be to build multiple classifiers to identify the individual attributes α_1 and α_2 independently, and to then identify locations where α_1 and α_2 co-exist within spatial proximity of each other. Unfortunately this approach does not apply to complex imagery where α_1 and α_2 are more than simply within spatial proximity of each other; they may in fact be spatially connected to each other in a specific fashion. Thus, there is a need for incorporating domain knowledge to link α_1 and α_2 so that the presence of object O can be identified. In this work, we exploit domain knowledge by leveraging the presence of lumen (α_1) surrounded by multiple nuclei (α_2) to identify a tubule O. We hypothesize that the spatial relationship between these two objects can be characterized by the O'Callaghan neighborhood [29], a specialized graph defined by distance and directional constraints, ideal for linking α_1 and α_2 in a domain contextual manner. Subsequently, image statistics quantifying the spatial arrangement of α_2 with respect to α_1 can be used to train a classifier to decide the presence (i.e. true lumen) or absence (i.e. false lumen) of an object O. We believe that true lumen (Figure 2.1(a), (b)) will be distinguishable from false lumen (Figure 2.1(c)) based on the proximity, order, and spacing of the nuclei in the O'Callaghan neighborhood for each luminal area.

2.8.3 Sampling and consensus classification strategy for whole-slide DP analysis

The third goal of this work is to develop an effective strategy for the sampling, feature extraction, and classification of whole-slide DP images. Due to the limitations of arbitrary, randomized, and hierarchical FOV selection, we hypothesize that a multi-field-of-view (multi-FOV) framework, which automatically integrates image features from multiple FOVs at various sizes, will lead to accurate classification of whole-slide DP images [2, 3, 7]. Clinicians implicitly incorporate multiple FOVs of different sizes during visual analysis; yet, the selection of an optimal FOV (i.e. image patch) size for computerized analysis of whole-slide DP slides is not straightforward. For example, in Figure 2.2(a), while the smallest FOV simply looks like necrotic tissue, the medium-sized FOV would be accurately classified as ductal carcinoma *in situ* (DCIS). At the other end of the spectrum, the largest FOV (i.e. entire image) containing both DCIS and invasive cancer would be classified ambiguously since it is too heterogeneous. It is important to note that a multi-FOV framework differs from traditional multi-scale (i.e. multi-resolution) classifiers that operate on a fixed FOV at multiple spatial resolutions [45, 79, 80] (Figure 2.2(b)). We believe the multi-FOV approach confers a number of advantages in both theory and practice, including the (1) integration of image information from a multitude of FOVs of different sizes and (2) potential for efficient parallelized implementation via concurrent analysis of different FOV sizes. In addition, we believe that the multi-FOV framework will be extensible to other types of DP (e.g.



(a)

(b)

Figure 2.2: (a) The multi-FOV framework presented in this paper operates by maintaining a fixed scale while analyzing several FOV sizes. (b) Conversely, a multi-scale framework would operate by analyzing a fixed FOV at different spatial resolutions.

immunohistochemically (IHC) stained slides), allowing for the integration of different types of prognostic information from multi-parametric histopathological studies.

Specifically, the main novel contributions are:

- A multi-field-of-view (multi-FOV) classifier able to apply a single operator across a multitude of fields of view at different sizes in order to extract relevant QH information,

- The incorporation of a robust feature selection scheme into the multi-FOV framework to independently identify salient image features at each FOV size,

- An image-based classifier that comprehensively analyzes whole-slide DP images rather than arbitrarily or randomly selected FOVs, and

- A multi-parametric extension to the multi-FOV framework that incorporates image-based features from other types of histological studies (e.g. IHC staining) to achieve an improved prediction of disease outcome.

The fourth goal of this work is to overcome the major constraints on classifier selection in decision support systems that employ medical imaging data. We address crucial questions that arise early in the development in a classification system, namely:

- Given a small pilot dataset, can we predict the error rates associated with a classifier assuming that a larger data cohort will become available in the future?
- Will the relative performance between multiple classifiers hold true as data cohorts grow larger?

Due to the heterogeneous nature of medical imaging data [24, 26], we believe that the traditional RRS-based approach originally used to model gene microarray data [82] will be inadequate for modeling classifier performance. This is exemplified in Figure 2.3 by the variability exhibited by the calculated (black boxes) and extrapolated (blue curves) error rates resulting from the use of different training and testing pools from the same dataset. However, we hypothesize that the RRS framework can be extended for robust application to medical imaging data. The specific novel contributions of this work are:

- More stable learning curves by the incorporation of cross-validation into the RRS scheme, which ensures that all samples are used at least once for both classifier training and testing,
- A direct comparison of performance across multiple classifiers as dataset size increases, and
- Enabling a power analysis of classifiers operating on the pixel level (as opposed to patient/sample level), which cannot be currently done via standard sample power calculators.

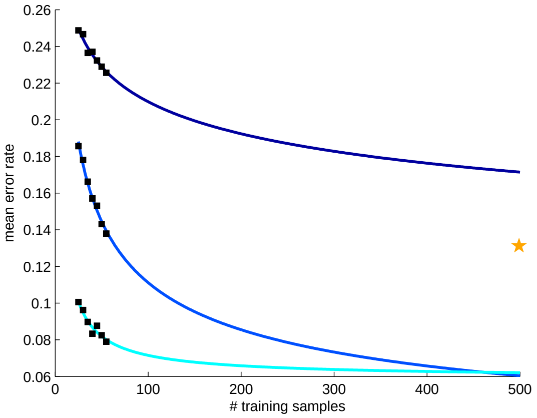


Figure 2.3: Application of traditional repeated random sampling (RRS) to heterogeneous medical data yields highly variable calculated (black boxes) and extrapolated (blue curves) mean error rates. Each set of error rates is derived from an independent RRS trial that employs different training and testing pools for the classification of cancerous and non-cancerous prostate cancer histopathology via a naive Bayes classifier. The yellow star represents the leave-one-out cross-validation error (i.e. the expected lower bound on error) produced by a larger validation cohort.

Chapter 3

EM-based segmentation-driven color standardization of DP images

3.1 Specific notation for this chapter

For all methods, an image scene $\mathcal{C}_a = (C, \mathbf{f})$ is a 2D set of pixels $c \in C$ and $\mathbf{f}$ is the associated function that assigns RGB values. In addition, the function g is used to assign intensity values (from the HSI color space) for evaluation purposes. A subscene $\mathcal{D}_a \subset \mathcal{C}_a$ is defined as a portion of the image scene $\mathcal{C}_a$ as identified by the Expectation-Maximization (EM) algorithm, which is used to partition $\mathcal{C}_a$ into κ components. Additional notation and symbols are defined in Appendix A.

3.2 EM-based Partitioning of Broad Tissue Classes

The EM framework is employed to first partition each image into broad tissue classes. First, the pixels in each image scene $\mathcal{C}_a$ are modeled as a Gaussian mixture of κ components, where $K \in \{1, 2, \ldots, \kappa\}$. We optimize the model parameter set $\gamma^i = \{\mu_K^i, \sigma_K^i, \mathbf{p}_K^i : \forall K\}$, comprising the mean μ_K, covariance σ_K, and *a priori* probability $\mathbf{p}_K$ at iteration i. The mixture is initialized to γ^0 via κ-means clustering of RGB values over all $c \in C$. The Expectation step calculates the posterior probability

$$p^i(K|\mathbf{f}(c)) = \frac{\mathbf{p}_K^i N(\mathbf{f}(c)|\mu_K^i, \sigma_K^i)}{\sum_{j=1}^{\kappa} \mathbf{p}_K^i N(\mathbf{f}(c)|\mu_j^i, \sigma_j^i)},$$

where $N(\mathbf{f}(c)|\mu_K, \sigma_K)$ represents the value of Gaussian component K at RGB value $\mathbf{f}(c)$. The Maximization step uses p^i to calculate $\gamma^{i+1} = \{\mu_K^{i+1}, \sigma_K^{i+1}, \mathbf{p}_K^{i+1}\}$ [87].

$$\mu_K^{i+1} = \frac{\sum_{c \in C} p(K|\mathbf{f})\mathbf{f}}{\sum_{c \in C} p(K|\mathbf{f})}$$

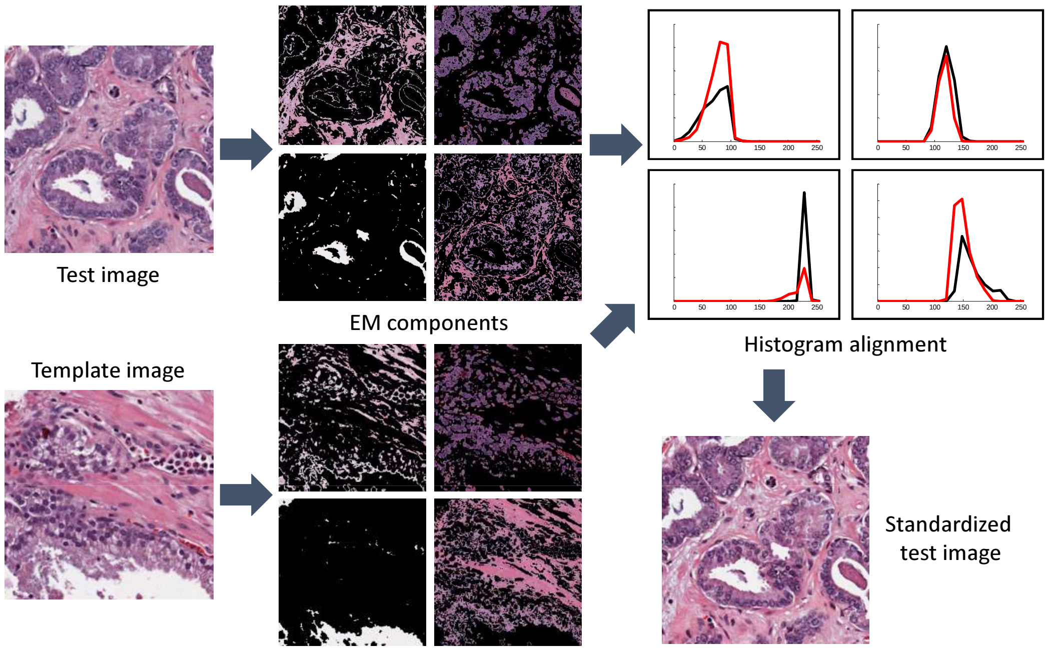


Figure 3.1: EM-based standardization (EMS) first decomposes the test and template images into a pre-determined number of components via the Expectation Maximization algorithm. The distribution for each component in the test image is aligned independently to the corresponding component in the template image. The standardized components are subsequently recombined to create a standardized test image.

$$\sigma_K^{i+1} = \frac{\sum_{c \in C} p(K|\mathbf{f})(\mathbf{f} - \mu_K^i)(\mathbf{f} - \mu_K^i)^\mathsf{T}}{\sum_{c \in C} p(K|\mathbf{f})}$$

$$\mathbf{p}_K^{i+1} = \frac{1}{|C|} \sum_{c \in C} p(K|\mathbf{f})$$

The EM algorithm converges when $\|(\mathcal{L}^{i+1} - \mathcal{L}^i)/\mathcal{L}^i\| < \epsilon$, where $\mathcal{L}^i$ is the log likelihood of the Gaussian mixture model with parameters γ^i and $\epsilon = 10^{-5}$ determined empirically. The appropriate class $K^* = \text{argmax}_K \, p(K|\mathbf{f}(c))$ is found for each pixel $c \in C$ by identifying the maximum posterior probability over all $K \in \{1, 2, \ldots, \kappa\}$. Hence, we are able to define a subset of pixels $D_K \subset C$ corresponding to tissue class K from image scene $\mathcal{C}_a$.

3.3 Determining Correspondence of Tissue Classes from Different Images

Since the EM algorithm performs tissue partitioning in an unsupervised manner, correspondence between tissue classes is not guaranteed across different images. For instance, the background (i.e. white) regions may coincide with the first EM component in one image and the final component in a second image. In this work, the identification of a corresponding tissue class between two images is performed automatically. Let mean RGB values $\mu_{K,a}$ and $\mu_{K,b}$ be defined for tissue class K in image scenes $\mathcal{C}_a$ and $\mathcal{C}_b$, respectively. The first pair of matching tissue classes is identified by minimizing the pairwise Euclidean distances between mean RGB values such that

$$\underset{i,j \in \{1,2,\ldots,\kappa\}}{\operatorname{argmin}} \|\mu_{i,a} - \mu_{j,b}\|.$$

The matching tissue classes are set aside and this process is subsequently repeated κ times until all tissue classes have been matched.

3.4 EM-based Segmentation-Driven Color Standardization (EMS) for Digital Pathology Images

We first describe the generalized scheme for color standardization using landmark-based piecewise linear interpolation [42] followed by an explanation of the EMS approach, which aligns color distributions of the broad tissue classes identified by the EM algorithm in Section 3.2.

3.4.1 Generalized Color Standardization

Let $\mathcal{D}_a \subset \mathcal{C}_a$ and $\mathcal{D}_b \subset \mathcal{C}_b$ correspond to sub-scenes in test image $\mathcal{C}_a$ and template image $\mathcal{C}_b$. For a single color channel (i.e. red, green, or blue), landmarks $\{r_{\min}, r_{10}, r_{20}, \ldots, r_{90}, r_{\max}\}$ and $\{s_{\min}, s_{10}, s_{20}, \ldots, s_{90}, s_{\max}\}$ are defined at the minimum and maximum, as well as evenly-spaced percentiles $\{10, 20, \ldots, 90\}$, of all pixel values in $\mathcal{D}_a$ and $\mathcal{D}_b$, respectively. Pixel values from the test image in the range $[r_{\min}, r_{10}]$ are interpolated to

Algorithm 1 GeneralizedStandardization()

Input: Template image $\mathcal{D}_b$. Test image $\mathcal{D}_a$ to be standardized.
Output: Standardized image $\hat{\mathcal{D}}_a$.

```
for RGB channels $i \in \{R, G, B\}$ in $\mathcal{D}_a$ and $\mathcal{D}_b$ do
    Define $\{r_{\min}, r_{10}, r_{20}, \ldots, r_{90}, r_{\max}\}$ as landmarks in $\mathcal{D}_a$.
    Define $\{s_{\min}, s_{10}, s_{20}, \ldots, s_{90}, s_{\max}\}$ as landmarks in $\mathcal{D}_b$.
    Interpolate pixel values from range $[r_{min}, r_{10}]$ to range $[s_{min}, s_{10}]$. Repeat for all sets of adjacent landmarks.
end for
Recombine standardized RGB channels to construct standardized image $\hat{\mathcal{D}}_a$.
```

match the corresponding landmarks $[s_{\min}, s_{10}]$ in the template image. After repeating this process for all sets of adjacent landmarks in all color channels, the standardized pixels are recombined to construct a standardized test scene $\hat{\mathcal{D}}_a$.

3.4.2 Class-Specific Color Standardization of Broad Tissue Classes

Algorithm 2 shows how EMS extends the generalized standardization approach by incorporating prior domain knowledge of tissue composition in DP. First, the EM algorithm is applied to partition each image into κ tissue classes (Section 3.2) and corresponding tissue classes between test and template images are automatically matched (Section 3.3). For each tissue class, pixels from the test and template images are standardized using the piecewise linear interpolation method presented in Algorithm 1. Subsequently, standardized pixels from all tissue classes are recombined to create a standardized test image.

3.5 Experimental Design

3.5.1 Data Cohort

The EM-based color standardization scheme is evaluated on digitized H & E stained histopathology images from independent prostate (N=19) and oropharyngeal (N=26) cohorts, in which each image was taken from a different patient (Table 3.1). All images were digitized via a whole slide scanner at a spatial resolution of 1 µm/pixel and cropped to be 500 × 500 pixels in size. Both cohorts were empirically determined to have $\kappa = 4$

Algorithm 2 EMbasedStandardization()

Input: Template image $\mathcal{C}_b$. Test image $\mathcal{C}_a$ to be standardized. Number of EM components κ.

Output: Standardized image $\mathcal{C}'_a$.

```
Apply EM algorithm to partition pixels from both $\mathcal{C}_a$ and $\mathcal{C}_b$ into $\kappa$ tissue classes.
Pair matching tissue classes between $\mathcal{C}_a$ and $\mathcal{C}_b$.
for $K \in \{1, 2, \ldots, \kappa\}$ do
    Define sub-scenes $\mathcal{D}_a^K \subset \mathcal{C}_a$ and $\mathcal{D}_b^K \subset \mathcal{C}_b$ corresponding to EM component $K$.
    Perform GeneralizedStandardization() using $\mathcal{D}_a^K$ and $\mathcal{D}_b^K$ as test and template images, respectively (Alg. 1).
end for
Create standardized image $\mathcal{C}'_a = \{\mathcal{C}_a^K : \forall K \in \{1, 2, \ldots, \kappa\}\}$ by recombining standardized sub-scenes from all $\kappa$ components of the test image.
```

Cohort	# images	Staining	Resolution	Size
Prostate	19	H & E	1 µm/pixel	500 × 500 pixels
Oropharyngeal	26			

Table 3.1: A description of the prostate and oropharyngeal data cohorts used in this chapter.

broad tissue classes corresponding to nuclei, epithelium, stroma, and background (i.e. white space). In addition, one image from each cohort is designated as a template image to which all other (test) images are aligned (Figure 3.2). It is important to note that a single "ideal" template image does not exist for all applications; instead, the template is selected based on its performance in terms of the desired task (e.g. nuclear segmentation).

3.5.2 Comparative Strategy: Global Standardization

In addition to the comparison against unstandardized images, the ability of EMS to align color distributions is evaluated against global standardization (GS). GS is a straightforward approach to color standardization that does not account for the heterogeneous tissue structure in DP images. Instead of partitioning each image into multiple tissue classes, a single histogram is constructed from all pixels in a test image and aligned to the entire histogram from the template image. Specifically, the GS approach can be considered a modified application of GeneralizedStandardization() from Algorithm 1 using entire image scenes $\mathcal{C}_a$ and $\mathcal{C}_b$ used as the test and template images, respectively.

3.5.3 Performance Evaluation Measures

Qualitative Segmentation Consistency Across Images

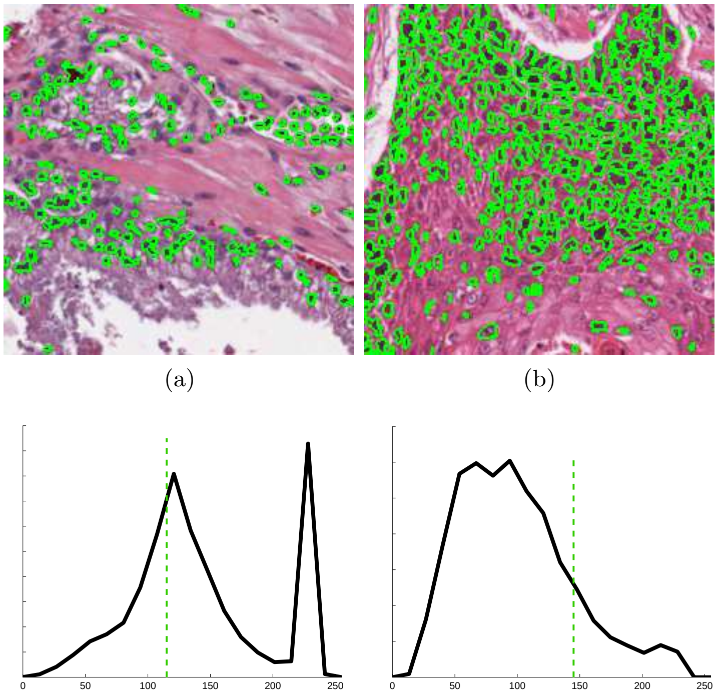
A qualitative evaluation of color standardization in DP images is performed by observing the consistency of tissue segmentation across several images in a cohort [43]. For image $\mathcal{C}_a$, we segment pixels corresponding to nuclei $D = \{c : c \in C, g(c) \in [0, \psi]\}$, where ψ is a threshold in the intensity channel g from the hue-saturation-intensity (HSI) color space. Given intensities that occupy the range $g(c) \in [0, 255]$, thresholds of $\psi = 115$ for the prostate cohort and $\psi = 145$ for the oropharyngeal cohort were selected empirically for their ability to provide a basic nuclear segmentation in their respective template images (Figure 3.2). Visually, images from a standardized cohort should yield a more consistent segmentation of nuclei (in comparison to the template image) than the original set of unstandardized images.

Quantitative Evaluation via Normalized Median Intensity

The segmentation results from Section 3.5.3 can also be evaluated quantitatively via normalized median intensity (NMI), which is employed to characterize color constancy from a segmented region across all images in a dataset. Using the segmented nuclear region $D \in C$ for an image $\mathcal{C}_a$, NMI is defined as

$$\text{median}\left(\frac{g(D)}{\max g(D)}\right),$$

where $g(D)$ is the set of intensities for all pixels isolated by the thresholding process. Intensity values across all images in a cohort are considered to be more consistent as (1) the standard deviation and (2) the coefficient of variation (i.e. standard deviation divided by mean) of NMI decreases.



(a) (b)

(c) (d)

Figure 3.2: The template images selected for (a) prostate and (b) oropharyngeal cohorts are shown along with segmented nuclei (green outline). (c), (d) Corresponding green color channel distributions are shown along with a dotted green line denoting the location of the empirically-selected threshold used to segment the nuclei.

Quantitative Evaluation via Histogram Landmark Distance

Another measure of improved standardization is the distance between corresponding histogram landmarks (i.e. percentiles $\{10, 20, \ldots, 90\}$) in the template and test images (Figure 3.3), whereby histograms are considered to have improved alignment as the mean landmark distance decreases. Using the notation defined in Algorithm 1, the mean histogram landmark distance between a test image $\mathcal{C}_a$ and template image $\mathcal{C}_b$ can be defined as $\phi(a, b) = \frac{1}{9} \sum_{j \in \{10,\ldots,90\}} \|r_j - s_j\|$, where r_j and s_j are corresponding landmarks in $\mathcal{C}_a$ and $\mathcal{C}_b$, respectively. For a set of H histograms, pairwise landmark distance $\Phi = \{\phi(a, b) : \forall a, b \in \{1, \ldots, H\}, a \neq b\}$ is calculated between all histograms and mean distance is reported for the cohort.

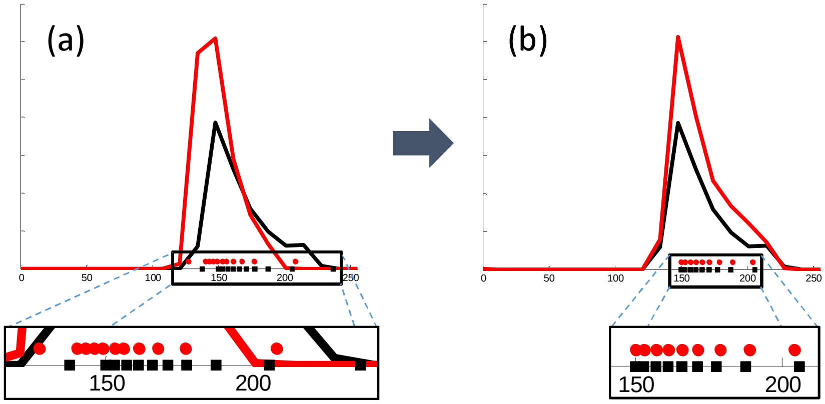


Figure 3.3: Histograms representing a single EM component are shown for template (black) and test (red) images both (a) before standardization and (b) after EMS has been applied. The insets illustrate the evenly-spaced percentiles $\{10, 20, \ldots, 90\}$ used as landmarks for histogram alignment during the standardization process.

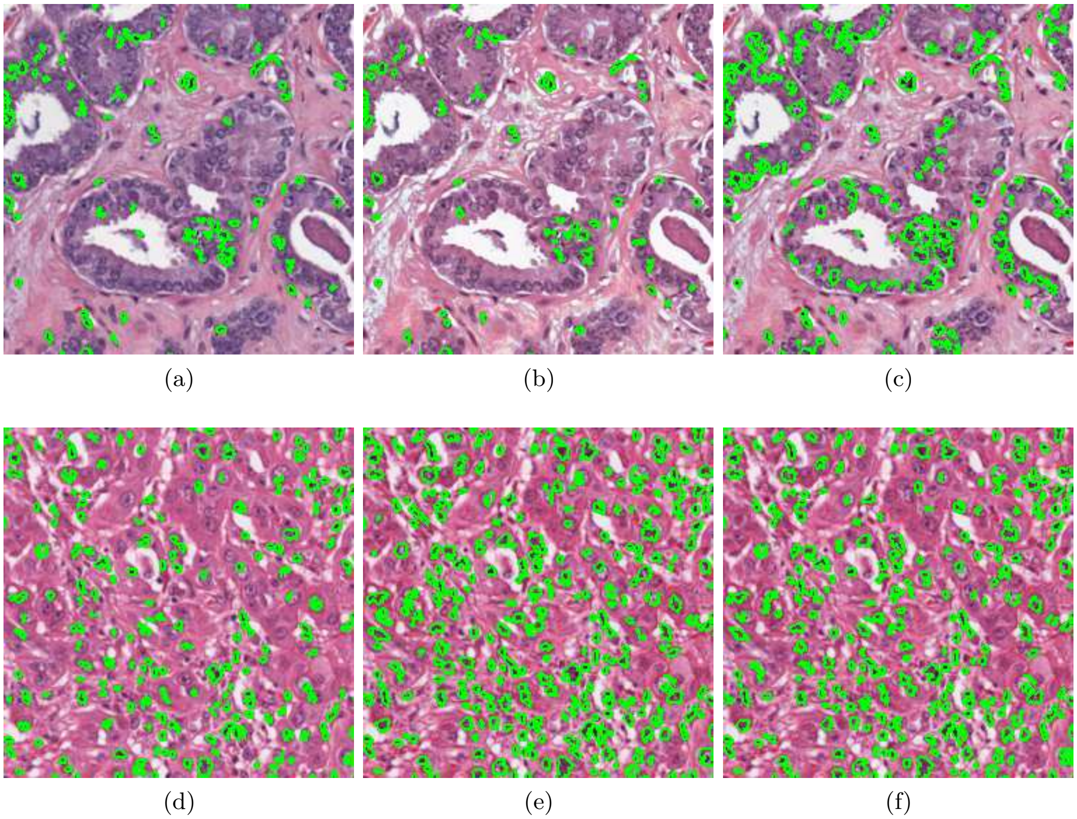
3.6 Results and Discussion

3.6.1 Qualitative Evaluation of Consistency in Nuclear Segmentation

Qualitative evaluation is performed by visualizing the effect of standardization on segmentation of nuclei in the test images (Figure 3.4). The inconsistent segmentation results between the template images (Figures 3.2(a), (b)) and unstandardized test images (Figures 3.4(a), (d)) clearly demonstrates the inherent color nonstandardness that affects DP images. A more consistent segmentation of nuclear regions is visible after GS (Figures 3.4(b), (e)) and is further improved by the application of EMS (Figures 3.4(c)-(f)). The improvement seen by employing EMS suggests that separation of tissue classes may be vital to the development of algorithms for the segmentation of primitives (e.g. nuclei).

3.6.2 Quantitative Evaluation of Segmentation Consistency via NMI

The qualitative results presented in Figure 3.4 are also evaluated quantitatively by calculating the normalized median intensity (NMI) of the segmented regions [42]. In terms of NMI, EMS produces improved color constancy compared to the original images, with considerably lower NMI standard deviation (SD) of 0.0054 vs. 0.0338 and NMI coefficient of variation (CV) of 0.0062 vs. 0.0393 in the prostate cohort (Table 3.2). In



(a) (b) (c)

(d) (e) (f)

Figure 3.4: H & E stained test images for (a)-(c) prostate and (d)-(f) oropharyngeal cancers are shown. Segmented nuclei (green outline) are shown for images that are (a), (d) unstandardized (b), (e) globally standardized (GS), and (c), (f) standardized via EMS.

addition, EMS yields more consistent results than GS, demonstrating an order of magnitude improvement in SD and CV of 0.0305 and 0.0354, respectively. All corresponding results for the oropharyngeal cohort show similar improvement after standardization.

3.6.3 Quantitative Evaluation of Histogram Landmark Alignment

The performance of EMS is further supported by histograms of the green color channel (from the RGB color space) for both prostate and oropharyngeal cohorts (Figure 3.5 and 3.6). Examining the prostate cohort, it is visually clear that unstandardized images have highly misaligned color distributions for both the global histogram (Figure 3.5(a)) and for the EM component corresponding to nuclei (Figures 3.5(d)). While both GS and EMS yield improved alignment over unstandardized images, a closer examination

	Prostate		Oropharyngeal	
	SD	CV	SD	CV
Original	0.0338	0.0393	0.0261	0.0302
GS	0.0305	0.0354	0.0166	0.0193
EMS	**0.0054**	**0.0062**	**0.0034**	**0.0039**

Table 3.2: Standard deviation (SD) and coefficient of variation (CV) of normalized median intensity (NMI) is calculated across all images in the prostate and oropharyngeal cohorts.

of the GS distributions suggests that higher pixel values (denoted by the black dashed rectangle in Figure 3.5(b)) frequently suffer from poor alignment. This is because GS is unable to account for the large variations in the amount of white space (e.g. luminal areas, adipose tissue, slide background) across histopathology images. By contrast, EMS does not suffer from this issue (Figures 3.5(c)) since it partitions the white regions in each image and aligns their distributions independently. The histograms in Figure 3.6 suggest that similar results and trends hold true for the oropharyngeal cohort.

The improved alignment of EMS distributions over both unstandardized and GS distributions is confirmed quantitatively by calculating the histogram landmark distance for individual EM components (Figures 3.5(d)-(f)). Using the EM component corresponding to nuclei in the prostate histopathology images, EMS yields a significantly lower mean landmark distance of 2.25, compared to 54.8 and 27.1 for unstandardized and GS distributions, respectively (Figures 3.5(d)-(f)). The significance of this comparison is verified by application of the non-parametric Wilcoxon rank-sum test [88] in conjunction with a null hypothesis that pairwise histogram landmark distances between unstandarized images are not different from the distances calculated from standardized images (Table 3.3). Similarly, oropharyngeal images (Figures 3.6(d)-(f)) demonstrate significantly lower distances for EMS (4.2) compared to unstandardized (27.3) and GS (8.8) distributions, suggesting that EMS is able to more accurately account for the different proportions of tissue types (e.g. nuclei, epithelium, stroma, lumen) present in histopathology imagery.

Cohort	Unstandardized vs. GS	Unstandardized vs. EMS	GS vs. EMS
Prostate	< 0.0001	< 0.0001	< 0.0001
Oropharyngeal	0.0645	< 0.0001	< 0.0001

Table 3.3: P-values from application of the Wilcoxon rank-sum test to pairwise histogram landmark distances between all images in the prostate and orophryngeal cohorts. All reported p-values have undergone Bonferroni correction for multiple comparisons [89].

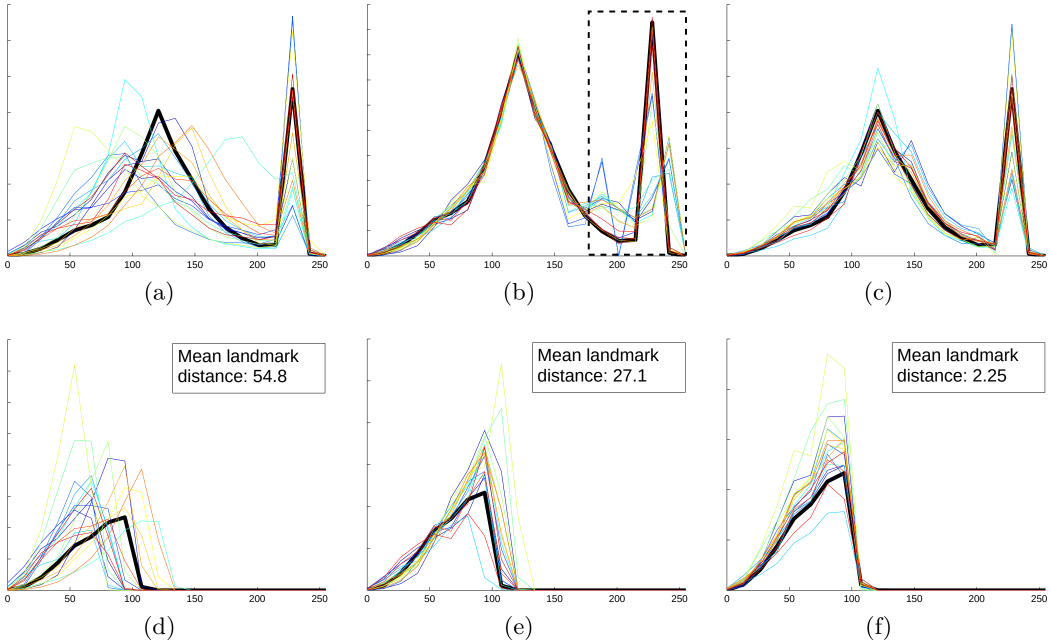
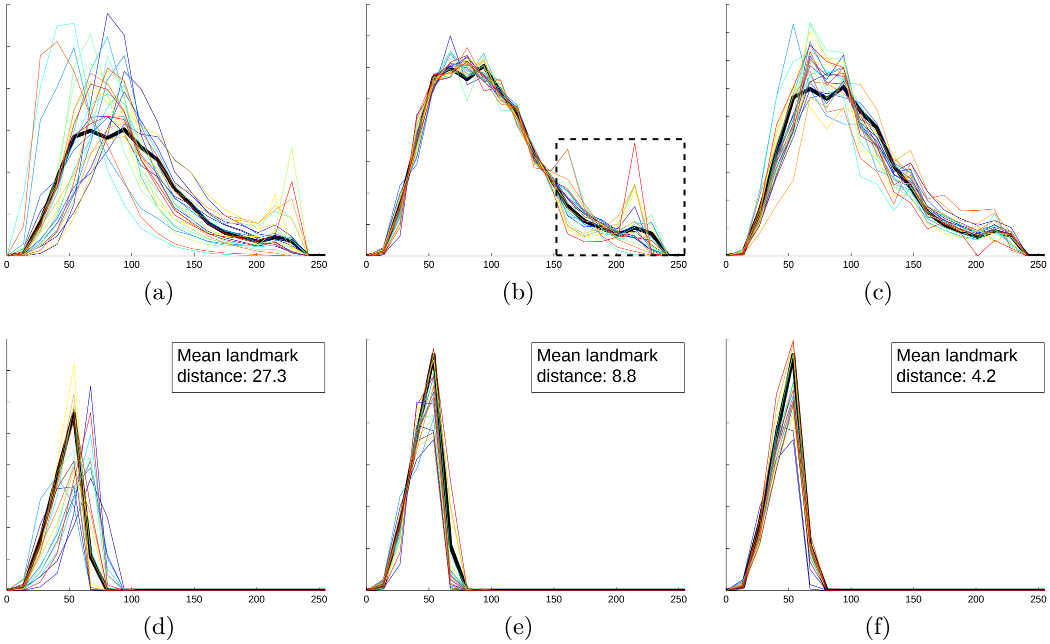


Figure 3.5: Distributions of the green color channels are shown for all images in the prostate cohort. Results are shown for (a), (d) unstandardized, (b), (e) GS, and (c), (f) EMS images. Alignment is shown for histograms representing (a)-(c) entire images and (d)-(f) an individual EM component (i.e. tissue class) along with the mean pairwise landmark distance over all images. In each figure, the histogram of the template image is represented by a thick black line and misalignment associated with GS is highlighted by a black box with dashed line.



(a) (b) (c)

(d) (e) (f)

Figure 3.6: Distributions of the green color channels are shown for all images in the oropharyngeal cohort. Results are shown for (a), (d) unstandardized, (b), (e) GS, and (c), (f) EMS images. Alignment is shown for histograms representing (a)-(c) entire images and (d)-(f) an individual EM component (i.e. tissue class) along with the mean pairwise landmark distance over all images. In each figure, the histogram of the template image is represented by a thick black line and misalignment associated with GS is highlighted by a black box with dashed line.

Chapter 4

Detection and segmentation of clinically relevant tissue structures in breast cancer DP

4.1 Specific notation for this chapter

Given an image scene $\mathcal{C} = (C, \mathbf{g})$ comprised of a 2D pixel grid C and vectorial function $\mathbf{g}$ assigning the RGB color space, let pixels $o^n \in C$ and $o^\ell \in C$ correspond to the centroids of nuclei and potential lumen area, respectively. Similarly, $\mathbf{N} = \{o_1^n, o_2^n, \ldots, o_N^n\}$ and $\mathbf{L} = \{o_1^\ell, o_2^\ell, \ldots, o_L^\ell\}$ are defined as the sets of all N nuclei and L potential lumen, respectively, in $\mathcal{C}$. In addition, we define parameters for the distance T_r and directional T_θ constraints of the O'Callaghan neighborhood. Other commonly used notation can be found in Appendix A.

4.2 Isolating the hematoxylin stain using color deconvolution

Color deconvolution [90] is used to convert an image from the RGB color space $\mathbf{g}$ to a new color space $\bar{\mathbf{g}}$ defined by hematoxylin H, eosin E, and background K (i.e. white) channels (Figures 4.1(b), (c)). The relationship between color spaces $\mathbf{g}$ and $\bar{\mathbf{g}}$ is defined as $\mathbf{g} = \mathbf{A}\bar{\mathbf{g}}$, where the transformation matrix is given by

$$\mathbf{A} = \begin{bmatrix} \hat{H}_R & \hat{H}_G & \hat{H}_B \\ \hat{E}_R & \hat{E}_G & \hat{E}_B \\ \hat{K}_R & \hat{K}_G & \hat{K}_B \end{bmatrix}, \tag{4.1}$$

where $\hat{H}_R$, $\hat{H}_G$, and $\hat{H}_B$ denote the pre-defined, normalized red, green, and blue values, respectively, for the H channel. The second and third rows of $\mathbf{A}$ are defined analogously for the E and K channels, respectively. In this work, the pre-defined values in $\mathbf{A}$ are selected based on published values by Ruifrok and Johnston [90]. The intensity of a pixel c in the new color space is defined as $\bar{\mathbf{g}}(c) = \mathbf{A}^{-1}(c)\mathbf{g}(c)$, where $\mathbf{g}(c)$ and $\bar{\mathbf{g}}(c)$ are 3×1 column vectors. The extent of hematoxylin staining in image scene $\mathcal{C}$ (i.e. the hematoxylin channel) is subsequently defined as $H(C) = \{H(c) : \forall c \in C\}$ (Figure 4.1(d)).

4.3 Detection of nuclei in H & E stained DP using color deconvolution

Centroids of individual nuclei are identified by applying morphological opening to the hematoxylin channel identified in Section 4.2 and thresholding the result (Figures 4.1(e)-(g)). Note that this method does not detect each and every nucleus in an image. Previous work Ali et al. [63] as well additional work presented in Appendix C suggest that perfect identification of each and every nucleus may not be crucial for distinguishing patients with good and poor outcome. Hence, we present a high-throughput method that identifies a sufficient number of nuclei to reflect the clinically relevant variations in nuclear architecture.

4.4 Segmentation of nuclei in H & E stained DP using the color gradient based geodesic active contour

To segment nuclear regions, the hematoxylin channel identified in Section 4.2 is used to initialize a color gradient based geodesic active contour (CGAC) model developed by Xu et al. [91]. The CGAC approach represents an improvement over the traditional GAC model by employing a color gradient $\psi(\mathbf{g}(c)) = \frac{1}{1+s(\mathbf{g}(c))}$ for edge detection rather than the more common grayscale gradient [92]. The final boundaries of the CGAC segmentation are used to define a mask denoting nuclear regions (Figure 4.1(j)). Note that we aim to segment only nuclei belonging to epithelial cells while avoiding the darker nuclei representing lymphocytes and fibroblasts.

4.5 Segmentation of potential luminal areas in H & E stained DP using hierarchical normalized cuts initialized color gradient based geodesic active contour

Similar to the segmentation of nuclear regions performed in Section 4.4, the CGAC model [95] is used to segmented potential luminal (i.e. white colored) areas. However, in this case, a robust initialization of the CGAC model is provided by using the hierarchical normalized cuts (HNCut) algorithm to detect white areas within the image. The HNCut scheme [57] pyramidally traverses and reduces the color space of an image using a combination of the mean shift clustering [96] and normalized cuts [97] algorithms. This approach efficiently and accurately segments all potential lumen objects in the image, requiring minimal user interaction in the form of a color swatch (i.e. a few pixels) selected from the object of interest (i.e. white luminal area). HNCut is particularly well suited to identifying potential luminal areas since white areas in DP images do not suffer from variability; hence, a color swatch taken from one image can likely be used for all other images as well (Figures 4.5(a)-(d)).

4.6 Tubule detection in breast cancer using O'Callaghan neighborhoods

4.6.1 Construction of the O'Callaghan neighborhood

The O'Callaghan neighborhood is defined as the subset of epithelial nuclei most closely surrounding a potential lumen area. Formally, given a set of potential lumen $\mathbf{L}$ and epithelial nuclei $\mathbf{N}$, a neighborhood of nuclei $\mathbf{N}^{\ell} \subset \mathbf{N}$ is defined around each potential lumen centroid $o^{\ell} \in \mathbf{L}$. The construction of the O'Callaghan neighborhood for o^{ℓ} can be summarized by the following steps.

<u>Step 1</u>: Find the nucleus $o_1^n \in \mathbf{N}$ nearest to o^{ℓ} and include it in $\mathbf{N}^{\ell}$.

<u>Step 2</u>: Define the distance constraint T_r (Section 4.6.2) using o_1^n.

<u>Step 3</u>: Update direction constraint T_{θ} (Section 4.6.2) based on all nuclei in $\mathbf{N}^{\ell}$.

Step 4: Find the next nearest nucleus to o^ℓ and add it to $\mathbf{N}^\ell$ if it satisfies the constraints outlined in Steps 2 and 3.

Step 5: Repeat Steps 3 and 4 until all $o^n \in \mathbf{N}$ have been considered.

Step 6: Extract features describing the spatial arrangement of nuclei in $\mathbf{N}^\ell$ with respect to o^ℓ (Table 4.1).

4.6.2 Spatial constraints

Epithelial nuclei are added to an O'Callaghan neighborhood on the basis of two spatial constraints. First, a distance constraint ensures that only nuclei within close proximity to the potential lumen area are included. Instead of defining a fixed radius, the O'Callaghan neighborhood excludes distant nuclei based on a relative distance that is proportional (by a factor of T_r) to the distance between the potential lumen o^ℓ and the nearest cancer nucleus o^n_1 (Figure 4.3(a)). Formally, given the centroids for a potential lumen $o^\ell \in \mathbf{L}$ and its nearest neighboring nucleus $o^n_1 \in \mathbf{N}^\ell$, a nucleus $o^n_j \in \mathbf{N}$ will be included in the neighborhood $\mathbf{N}^\ell$ if

$$\frac{\|o^\ell - o^n_j\|}{\|o^\ell - o^n_1\|} \leq T_r, \tag{4.2}$$

where $\|\cdot\|$ represents the L2 norm and $j \in \{1, 2, \ldots, N\}$.

Second, a direction constraint ensures that the O'Callaghan neighborhood will be representative of the arrangement of nuclei in a tubule, i.e. only one nucleus in each direction will be considered. To determine whether a new nucleus should be added to the neighborhood, we need to ensure that it does not lie "behind" any of the nuclei already included in the neighborhood. Given potential lumen $o^\ell \in \mathbf{L}$ and a nucleus in its O'Callaghan neighborhood $o^n_i \in \mathbf{N}^\ell$, we say that nucleus $o^n_k \in \mathbf{N}$ is "behind" o^n_i if the angle θ_k between vectors $\overrightarrow{o^n_i o^\ell}$ and $\overrightarrow{o^n_i o^n_k}$ is less than the pre-defined threshold T_θ (Figure 4.3(b)). Formally, given centroids for a potential lumen $o^\ell \in \mathbf{L}$ and a nucleus $o^n_i \in \mathbf{N}^\ell$ within its neighborhood $\mathbf{N}^\ell$, the nuclear centroid $o^n_j \in \mathbf{N}$ will be included in

$\mathbf{N}^\ell$ if

$$\frac{\|o^\ell - o_i^n\|^2 + \|o_i^n - o_j^n\|^2 - \|o^\ell - o_j^n\|^2}{\|o^\ell - o_i^n\| \cdot \|o_i^n - o_j^n\|} < T_\theta, \tag{4.3}$$

where $i, j, k \in \{1, 2, \ldots, N\}$ and $i \neq j \neq k$.

4.6.3 Detection and segmentation of nuclear and luminal structures

To detect tubule formation in BCa histopathology, we must first find the constituent objects in the form of epithelial nuclei and potential lumen areas. Centroids of all N nuclei in $\mathcal{C}$ are identified via color deconvolution (as described in Section 4.3) and recorded as the set of pixels $\mathbf{N} = \{o_1^n, o_2^n, \ldots, o_N^n\}$. In Figure 4.4, H & E stained histopathology images are shown (Figures 4.4(a)-(d)) along with their respective hematoxylin channels (Figures 4.4(e)-(h)) and resulting nuclear centroids (Figures 4.4(i)-(m)).

Similarly, centroids of all L potential lumen regions are identified using HNCut-CGAC (as described in Section 4.5) and recorded as $\mathbf{L} = \{o_1^\ell, o_2^\ell, \ldots, o_L^\ell\}$. Figures 4.5(a)-(d) show the initialization achieved by the HNCut algorithm for four different BCa histopathology images. Further refinement by the CGAC model yields final segmentation boundaries (Figures 4.5(e)-(h)).

4.6.4 O'Callaghan neighborhood-based features for distinguishing potential lumen belonging to tubule and non-tubule structures

A total of 22 features are extracted to quantify the spatial arrangement of nuclei $\mathbf{N}^\ell$ around each potential lumen centroid o^ℓ (Table 4.1). Note that the number in parenthesis for the following subsection titles reflects the number of features in the feature class.

Number of nuclei in O'Callaghan neighborhood (1)

Potential lumen areas that do not belong to tubules often have fewer nuclear neighbors that fall within the O'Callaghan constraints. Thus, the number of O'Callaghan nuclear neighbors $|\mathbf{N}^\ell|$ is calculated as a feature value for each o^ℓ.

Feature #	Feature Name	Description
1	Number of nuclei	Number of nuclei in neighborhood
2-6	Distance to nuclei	Distance between each nuclei in neighborhood and the lumen centroid
7-10	Circular fit	Fit circles to nuclei and measure deviation of nuclei from edge circle
11-13	Angle between adjacent nuclei	Angle between two vectors connecting the lumen centroid to two adjacent nuclei in neighborhood
14-16	Distance between adjacent nuclei	Distance between adjacent nuclei in neighborhood
17-22	Elliptical fit	Fit ellipse to nuclei and measure evenness in spatial distribution of nuclei

Table 4.1: The 22 features used to quantify the O'Callaghan neighborhood for each potential lumen.

Distance between nuclei and lumen centroid (5)

To quantify the evenness in the distribution of nuclei about the lumen centroid o^ℓ, the Euclidean distance $d(o^\ell, o_i^n) = \|o^\ell - o_i^n\|$ is calculated between o^ℓ and each neighboring nucleus $o_i^n \in \mathbf{N}^\ell$. The set of distances for all $o_i^n \in \mathbf{N}^\ell$ is defined as

$$D(o^\ell) = \left\{ d(o^\ell, o_i^n) : \forall i \in \{1, 2, \dots, N\} \right\}. \tag{4.4}$$

The mean, standard deviation, disorder, maximum, and range of D yield five feature values for each o^ℓ.

Circular fit (4)

Since tubule formation is often characterized by the arrangement of nuclei in a circular pattern around a lumen area, we extract features to quantify the circularity of $\mathbf{N}^\ell$. First, a circle $\mathcal{O}(o^\ell, r)$ is constructed with center at lumen centroid o^ℓ and radius r. The Euclidean distance $F(o^\ell, o_i^n, \mathcal{O}) = \|d(o^\ell, o_i^n) - r\|$ is calculated between a nuclear centroid o_i^n and the constructed circle $\mathcal{O}$ with radius r. The set of distances for all

$o_i^n \in \mathbf{N}^\ell$ is defined as

$$\mathcal{F}(o^\ell, \mathcal{O}) = \left\{ F(o^\ell, o_i^n, \mathcal{O}) : \forall i \in \{1, 2, \ldots, N\} \right\}. \tag{4.5}$$

In this paper, the mean of $\mathcal{F}(o^\ell, \mathcal{O})$ is calculated as a feature value, where circles $\mathcal{O}(o^\ell, r)$ with radius $r \in \{\max(D), \min(D), \mathrm{mean}(D), \mathrm{median}(D)\}$ are constructed, yielding four features for each o^ℓ.

Angle between adjacent nuclei in neighborhood (3)

Another key property of tubules is that nuclei are arranged at regular intervals around the white lumen area, which can be quantified by examining the angles between adjacent nuclei in the tubule. Thus, for each potential lumen centroid o^ℓ, let $\overrightarrow{o^\ell o_i^n}$ be the vector from lumen centroid o^ℓ to neighboring nuclei o^n. We denote $\overrightarrow{o^\ell o_j^n}$ as the vector from o^ℓ to an adjacent neighboring nucleus o_j^n. The set of angles between adjacent nuclei $o_i^n,\ o_j^n \in \mathbf{N}^\ell$ is defined as

$$A = \left\{ \arccos\left(\frac{\overrightarrow{o^\ell o_i^n} \cdot \overrightarrow{o^\ell o_j^n}}{\|\overrightarrow{o^\ell o_i^n}\| \cdot \overrightarrow{o^\ell o_j^n}} \right) : \forall i, j \in \{1, 2, \ldots, N\}, i \neq j \right\} \tag{4.6}$$

The mean, standard deviation, and disorder of A are calculated to yield three feature values for each o^ℓ.

Distance between adjacent nuclei in neighborhood (3)

Another way to ensure that nuclei are arranged at regular intervals is by calculating the distances B between adjacent neighboring nuclei $o_i^n,\ o_j^n \in \mathbf{N}^\ell$, such that

$$B = \left\{ \|o_i^n - o_j^n\| : \forall i, j \in \{1, 2, \ldots, N\}, i \neq j \right\} \tag{4.7}$$

Since the magnitude and variation in these distances should be small for nuclei belonging to tubules, the mean, standard deviation, and disorder of B are calculated to yield three feature values for each o^ℓ.

Elliptical fit (6)

In the preparation of 2D planar histopathology slides, if a tubule is sectioned at an oblique angle, the resulting lumen and nuclei appear to form an elliptical pattern (Figure 4.6). This phenomenon is modeled by constructing an ellipse that best fits all nuclei in $\mathbf{N}^{\ell}$ using the method described in [98]. Since tubules are inherently symmetrical structures, it is reasonable to expect a similar number of nuclei on all sides of the lumen area. To this end, nuclei in the O'Callaghan neighborhood are separated into groups on either side of major axis $\mathbf{N}^{\ell+} \subset \mathbf{N}^{\ell}$ and $\mathbf{N}^{\ell-} \subset \mathbf{N}^{\ell}$ (Figure 4.6). The value $|\mathbf{N}^{\ell+}| - |\mathbf{N}^{\ell-}|$ is calculated as a feature to capture the balance of nuclear distribution on either side of the major axis. Five additional features are calculated, including the lengths of the major and minor axes as well as statistics calculated from the distances between nuclei and the elliptical fit.

4.6.5 Experimental design

Dataset

In this study a total of 1226 potential lumen from 105 images (from 14 patients) was considered. All samples were taken from H & E stained BCa histopathology images digitized at 20x optical magnification (0.5 µm/pixel). For each image, an expert pathologist provided ground truth annotations delineating locations of all tubules. A total of 22 O'Callaghan features (Section 4.6.4) were calculated to describe the spatial arrangement of cancer nuclei in $\mathbf{N}^{\ell}$.

Differentiating potential lumen belonging to tubules and non-tubules

At the individual tubule level, we evaluate the ability of the O'Callaghan features to classify each potential lumen as either a tubular lumen $\mathcal{Y}(o^{\ell}) = \omega_1$ or a non-tubular lumen $\mathcal{Y}(o^{\ell}) = \omega_2$. Over a set of 50 cross-validation trials, the mean ROC curve and the mean area under the ROC curve (AUC) were calculated. In addition, the mean and standard deviation of the classification accuracy (at the ROC operating point) are calculated over all trials.

K-fold cross-validation via the random forest classifier

In this work, randomized 3-fold cross-validation is used in conjunction with a random forest classifier [99] to evaluate the ability of the 22 O'Callaghan features to distinguish tubular and non-tubular lumen. The K-fold cross-validation scheme [41], commonly used to overcome the bias from arbitrary selection of training and testing samples, first randomly divides the dataset into K subsets. The samples in K-1 subsets are used for training a classifier, while those from the remaining subset are tested. This process is repeated K times while rotating the subsets to ensure that all samples are evaluated exactly once. The random forest is a meta-classifier that uses bootstrap sampling to aggregate a large number of independent C4.5 decision trees and achieve a stable classification result [99]. The output of each C4.5 decision tree is probabilistic, denoting the likelihood that a potential lumen is a tubule.

4.6.6 Results and discussion

The capability of our system to identify tubules is directly related to the identification of both tissue structures (i.e. epithelial nuclei and potential lumen). As demonstrated in Figures 4.4(i)-(k) and Figures 4.5(e), (g), and (h), the color deconvolution and HNCut-CGAC algorithms are able to quickly and accurately detect cancer nuclei and potential lumen areas. However, Figure 4.5(m) suggests that false positive errors (i.e. potential lumen incorrectly identified as tubules) occur when nuclei are not detected correctly. Similarly, Figure 4.5(j) illustrates the false negative errors (i.e. potential lumen incorrectly identified as non-tubules) that occur when HNCut-CGAC does not identify a potential lumen in the image.

The mean ROC curve resulting from 50 trials of cross-validation (Figure 4.7(a)) along with an associated AUC value of 0.91 ± 0.0027 suggest that the O'Callaghan features are able to accurately distinguish tubular lumen from non-tubular lumen. This is further confirmed by a classification accuracy of 0.86 ± 0.0039 and a positive predictive value of 0.89 ± 0.014 at the ROC operating point over all 50 cross-validation trials (Table 4.2).

Experiment	Accuracy	Positive Predictive Value
Tubule Detection	0.86 ± 0.0039	0.89 ± 0.014

Table 4.2: Mean and standard deviation of classification accuracy and positive predictive values over 50 cross-validation trials. The data cohort contained 1226 potential lumen areas and 22 features describing the O'Callaghan neighborhood around each potential lumen.

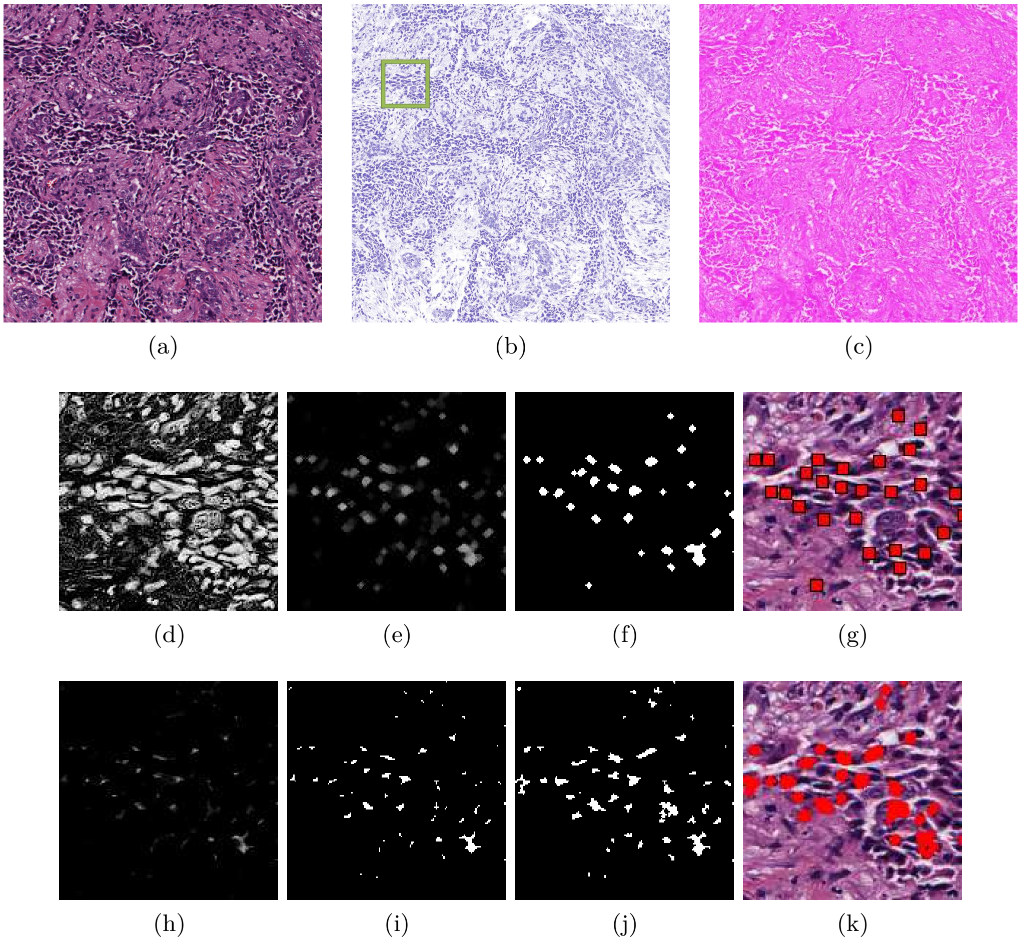


Figure 4.1: (a) A high grade histopathology image with its (b) hematoxylin and (c) eosin channels separated by color deconvolution. The green box in (b) denotes an inset providing more detailed visualization of the nuclear detection and segmentation process in (d)-(k). For nuclear detection, (d) the intensity of the hematoxlyin channel undergoes (e) morphological opening and (f) thresholding. (g) The centroids of the individual nuclei are later used for graph construction. The nuclear segmentation process also uses (d) the intensity of the hematoxylin channel, applying (h) morphological erosion, (i) thresholding, and (j) the color gradient based active contour model (CGAC), to achieve (k) a final segmentation result that is used for extraction of nuclear texture.

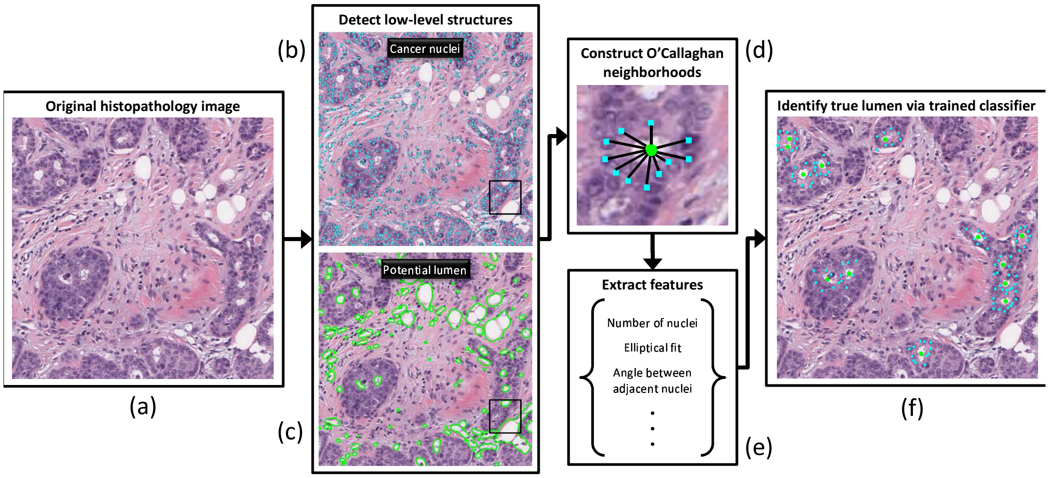

Original histopathology image
(a)
Detect low-level structures
(b)
Cancer nuclei
Potential lumen
(c)
Construct O'Callaghan neighborhoods
(d)
Extract features
Number of nuclei
Elliptical fit
Angle between adjacent nuclei
(e)
Identify true lumen via trained classifier
(f)

Figure 4.2: A flowchart detailing the methodological steps for our tubule detection system. Given (a) an original H & E stained histopathology image, low-level structures in the form of (b) epithelial nuclei and (c) potential lumen areas are first detected. (d) An O'Callaghan neighborhood is constructed around each potential lumen area and (e) image features are extracted to quantify the spatial linkage between the low-level structures. The features are then presented to (f) a trained classifier, which distinguishes true lumen areas (i.e. tubules) from false lumen areas (i.e. non-tubules).

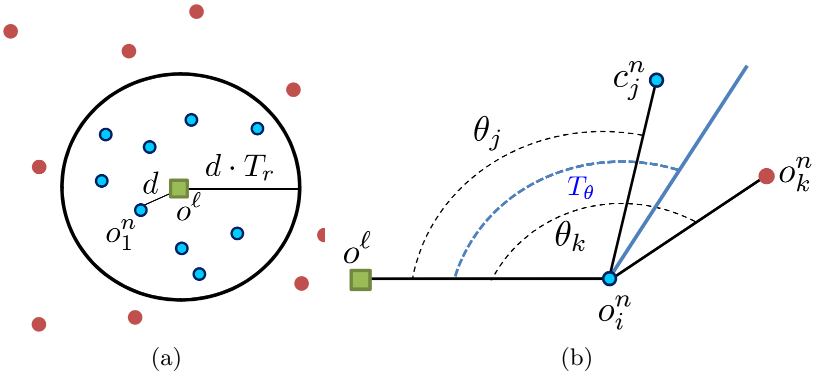


Figure 4.3: The O'Callaghan neighborhood is defined by both (a) distance and (b) direction constraints. In both schematics, the centroid of the potential lumen area o^ℓ (green squares), the centroids of nuclei that are disqualified by the constraints (red circles), and centroids of remaining nuclei that are still under consideration for inclusion in the neighborhood (blue circles) are illustrated. The distance constraint excludes nuclei outside a radius $d \cdot T_r$ based on the distance d between o^ℓ and nearest neighboring nucleus o_1^n. Given that nucleus o_i^n already included in the neighborhood, the direction constraint excludes nucleus o_k^n since angle $\theta_k > T_\theta$, where T_θ is a pre-defined threshold. Note that nucleus o_j^n may still be included since $\theta_j < T_\theta$.

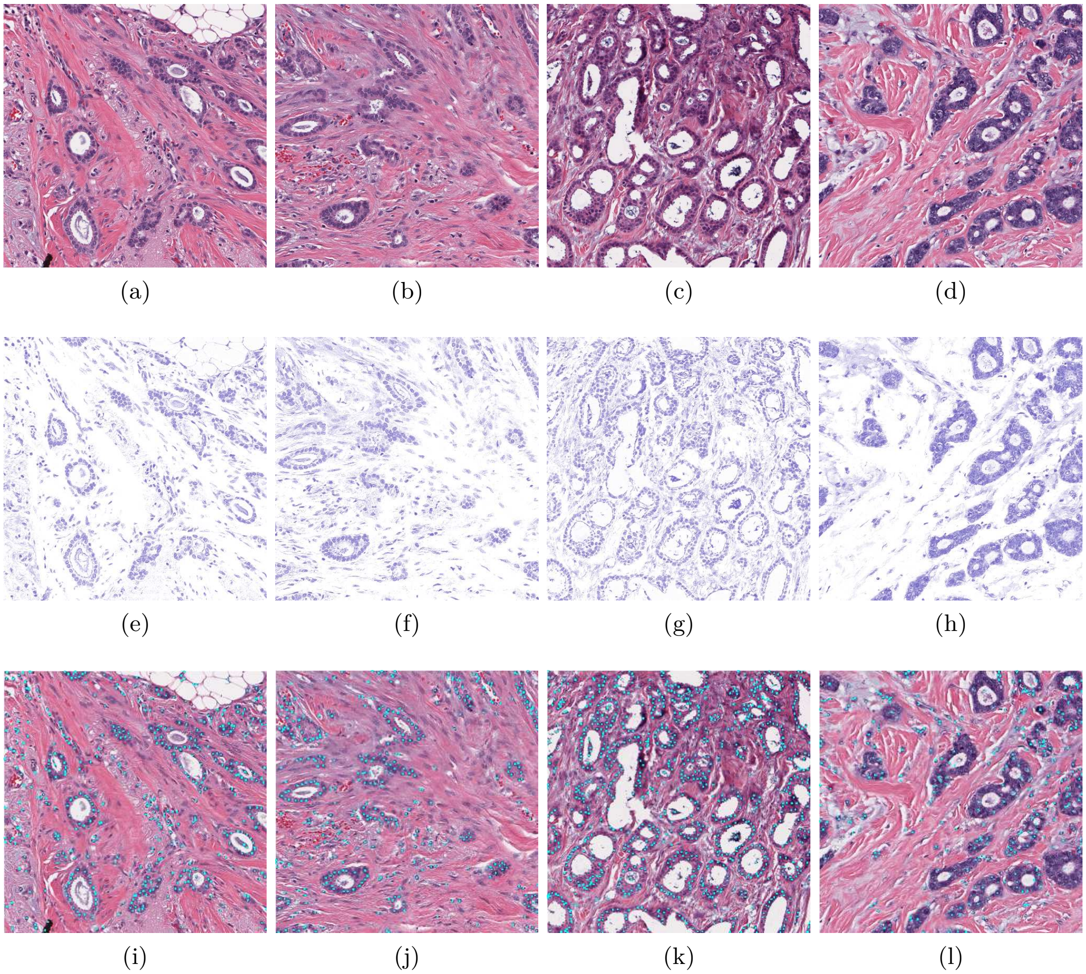
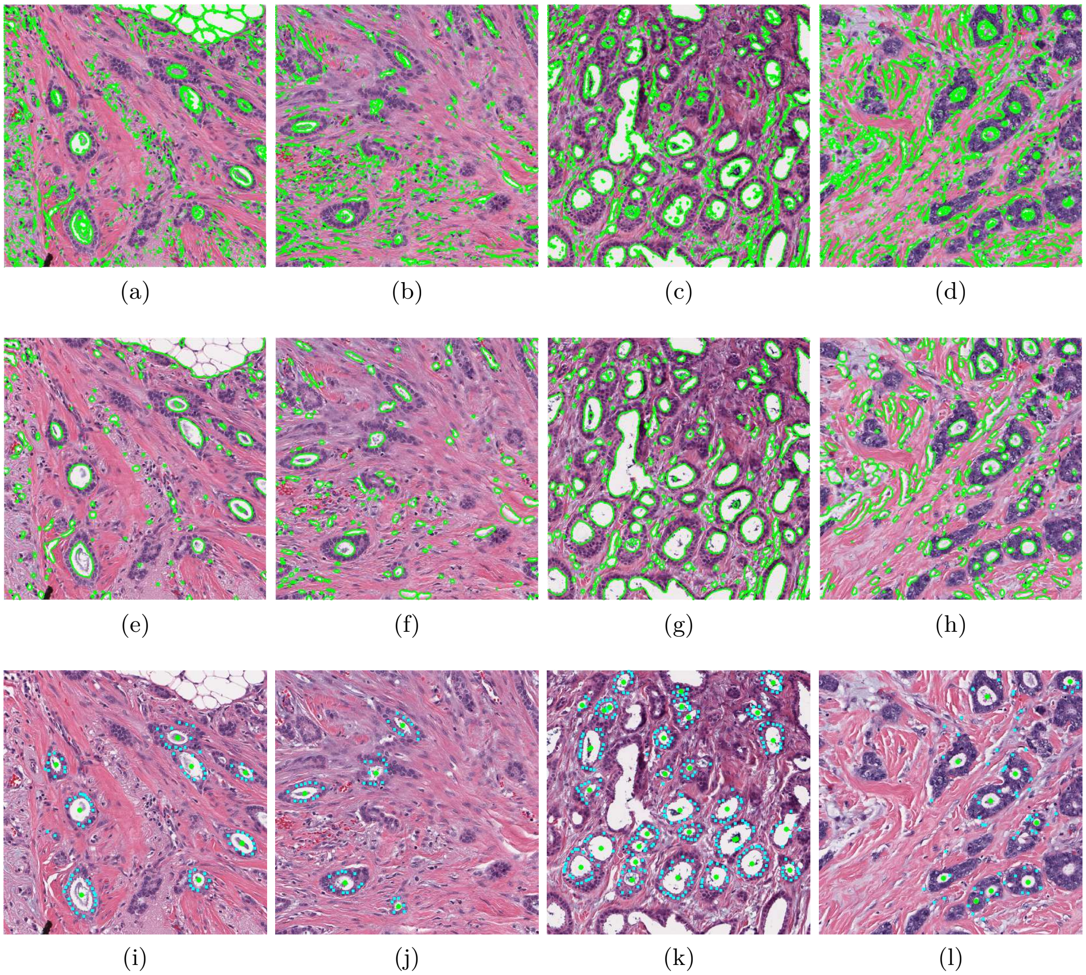


Figure 4.4: Automated nuclear detection is performed for (a)-(d) histopathology image patches by first using color deconvolution to isolate the corresponding (e)-(h) hematoxylin stain channel. Morphological opening is applied to the hematoxylin stain channel to isolate individual nuclear centroids ((i)-(m)).



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4.5: Segmentation of potential lumen is performed via two main steps. First, (a)-(d) a rough initial segmentation is achieved using the HNCut algorithm. This result is refined by the CGAC model and (e)-(h) a final segmentation is extracted. In (i)-(m), the centroids of only potential lumen classified as tubules (green circles) are shown along with the surrounding nuclei (blue squares) that comprise their respective O'Callaghan neighborhoods.

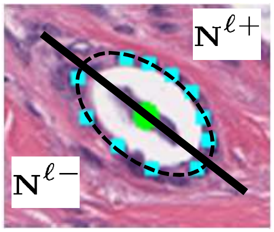


Figure 4.6: The centroid of a a potential lumen o^{ℓ} (green circle) is shown with the centroids of the nuclei in its O'Callaghan neighborhood $\mathbf{N}^{\ell}$ (blue squares). The ellipse (dashed black line) that best fits $\mathbf{N}^{\ell}$ is shown along with its major axis (solid black line). The nuclei on either side of the major axis are separated into the groups $\mathbf{N}^{\ell+}$ and $\mathbf{N}^{\ell-}$.

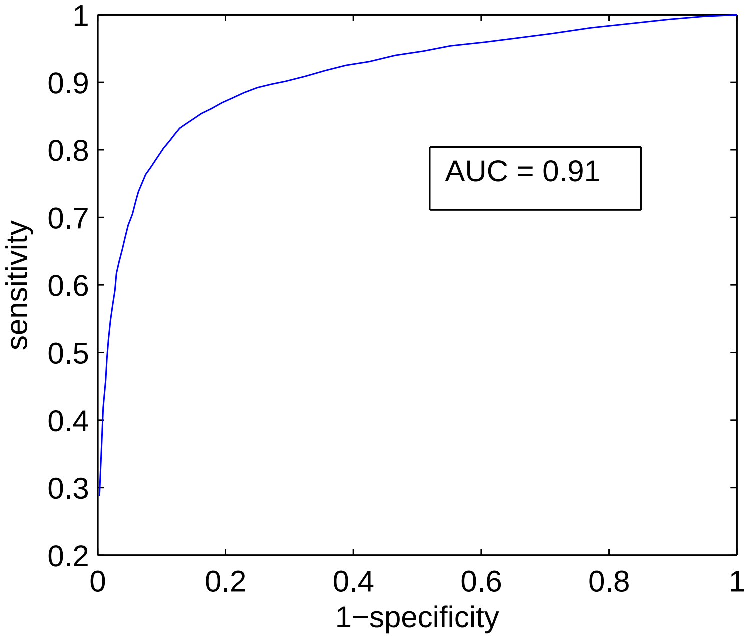


Figure 4.7: A mean ROC curve generated by averaging individual ROC curves from 50 trials of 3-fold cross-validation produces an AUC value of 0.91±0.0027 for differentiating potential lumen into tubular and non-tubular structures.

Chapter 5

Extraction of quantitative histomorphometric image features in breast cancer DP

This chapter describes the extraction of quantitative image features to characterize nuclear architecture $\mathbf{f}_{\mathrm{NA}}$, nuclear texture $\mathbf{f}_{\mathrm{NT}}$, and tubule density f_{TD} in H & E stained breast cancer DP images. In addition, we detail the extraction of vascular density f_{VD}, i.e. the quantification of microvessel formation, in CD34 IHC-stained DP images as a means of incorporating

5.1 Specific notation for this chapter

For image scene $\mathcal{C}$, we define $o^n \in C$ and $o^\ell \in C$ as centroids of nuclei and potential lumen areas, respectively. Similarly, $\mathbf{N} = \{o_1^n, o_2^n, \ldots, o_N^n\}$ and $\mathbf{L} = \{o_1^\ell, o_2^\ell, \ldots, o_L^\ell\}$ are defined as the sets of all N nuclei and L potential lumen, respectively, in $\mathcal{C}$. The subset of nuclei identified as belonging to true lumen is defined as $\hat{\mathbf{N}} \subset \mathbf{N}$. Other commonly used notation can be found in Appendix A.

5.2 Quantification of nuclear architecture via graph-based features

Utilizing individual nuclei as vertices for the construction of graphs allows for the quantification of tissue architecture. We define the complete, undirected graph $\mathcal{G} = (\mathbf{N}, \mathbf{E}, \mathbf{W})$, where $\mathbf{N} = \{o_1, o_2, \ldots, o_N\}$ is the set of vertices corresponding to the set of epithelial nuclear centroids, $\mathbf{E}$ is the set of edges connecting the nuclear centroids such that $\{(o_i, o_j) \in \mathbf{E} : \forall o_i, o_j \in \mathbf{N},\ i, j \in \{1, 2, \ldots, N\},\ i \neq j\}$, and $\mathbf{W}$ is a set of weights proportional to the length of each $E \in \mathbf{E}$. To extract information about the arrangement of nuclei, we construct subgraphs representing the Voronoi graph $\mathcal{G}_{\mathrm{VG}}$, Delaunay triangulation $\mathcal{G}_{\mathrm{DT}}$, and minimum spanning tree $\mathcal{G}_{\mathrm{MST}}$. In addition, statistics describing the number and density of nuclei are calculated directly from $\mathbf{N}$.

5.2.1 Voronoi graph

The Voronoi graph $\mathcal{G}_{\mathrm{VG}} = (\mathbf{N}, \mathbf{E}_{\mathrm{VG}}, \mathbf{W}_{\mathrm{VG}})$ (Figure 5.1(d)) is a spanning subgraph of $\mathcal{G}$ defined as a set of polygons $\mathbf{P} = \{P_1, P_2, \ldots, P_N\}$ surrounding all nuclear centroids $\mathbf{N}$ [100]. Each pixel $c \in C$ is linked with the nearest centroid $o \in \mathbf{N}$ (via Euclidean distance) and added to the associated polygon $P \in \mathbf{P}$. The mean, standard deviation, minimum/maximum (min/max) ratio, and disorder (i.e. standard deviation divided by the mean) are calculated for the area, perimeter length, and chord length over all $\mathbf{P}$, yielding a set of 13 features ($\mathbf{f}_{\mathrm{VG}}$) for each scene $\mathcal{C}$ (Table 5.1).

5.2.2 Delaunay triangulation

The Delaunay graph $\mathcal{G}_{\mathrm{DT}} = (\mathbf{N}, \mathbf{E}_{\mathrm{DT}}, \mathbf{W}_{\mathrm{DT}})$ (Figure 5.1(e)) is a spanning subgraph of $\mathcal{G}$ and the dual graph of $\mathcal{G}_{\mathrm{VG}}$ [100]. It is constructed such that if $P_i, P_j \in \mathbf{P}$ share a side, where $i, j \in \{1, 2, \ldots, N\}$, their nuclear centroids $o_i, o_j \in \mathbf{N}$ are connected by an edge $(o_i, o_j) \in \mathbf{E}_{\mathrm{DT}}$. The mean, standard deviation, min/max ratio, and disorder are calculated for the side length and area of all triangles in $\mathcal{G}_{\mathrm{DT}}$, yielding a set of 8 features ($\mathbf{f}_{\mathrm{DT}}$) for each scene $\mathcal{C}$ (Table 5.1).

5.2.3 Minimum spanning tree

A spanning tree $\mathcal{G}_{\mathrm{MST}} = (\mathbf{N}, \mathbf{E}_{\mathrm{MST}}, \mathbf{W}_{\mathrm{MST}})$ refers to any spanning subgraph of $\mathcal{G}$ [100]. The total weight $\widehat{\mathbf{W}}_{\mathrm{MST}}$ for each subgraph is determined by summing all individual weights $W \in \mathbf{W}_{\mathrm{MST}}$. The minimum spanning tree $\widehat{\mathcal{G}}_{\mathrm{MST}}$ (Figure 5.1(e)) is the spanning tree with the lowest total weight such that $\widehat{\mathcal{G}}_{\mathrm{MST}} = \mathrm{argmin}_{\mathcal{G}_{\mathrm{MST}} \in \mathcal{G}} \left[\widehat{\mathbf{W}}_{\mathrm{MST}} \right]$. The mean, standard deviation, min/max ratio, and disorder of the branch lengths in $\mathcal{G}_{\mathrm{MST}}$ yield a set of 4 features ($\mathbf{f}_{\mathrm{MST}}$) for each scene $\mathcal{C}$ (Table 5.1).

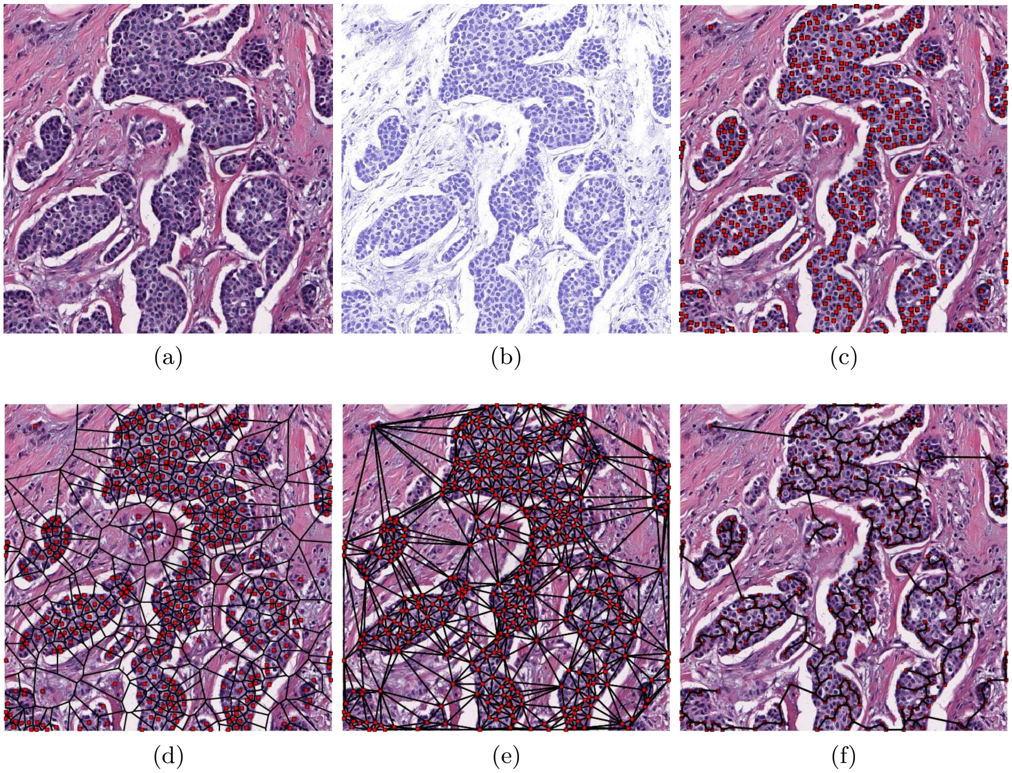


Figure 5.1: (a) Given an H & E stained image, (b) the (b) hematoxylin staining is isolated via color deconvolution and (c) thresholded to detect centroids of individual nuclei. The nuclei are used as vertices for the construction of (d) Voronoi, (e) Delaunay triangulation, and (f) minimum spanning tree graphs, from which 50 features describing nuclear architecture are extracted.

5.2.4 Nuclear Statistics

The global density $\frac{N}{|C|}$ of nuclei is calculated for each scene $\mathcal{C}$, where $|C|$ represents the number of pixels (cardinality) in $\mathcal{C}$. For any nuclear centroid $o_i \in \mathbf{N}$, we define a corresponding nuclear neighborhood $\eta^{\zeta}(o_i) = \{o_j : \|o_i - o_j\|_2 < \zeta, o_j \in \mathbf{N}, o_j \neq o_i\}$, where $\zeta \in \{10, 20, \ldots, 50\}$ and $\| \cdot \|_2$ is the L2 norm. The mean, standard deviation, and disorder of $\eta^{\zeta}(o_i), \forall o_i \in \mathbf{N}$ are calculated. Additionally we estimate the minimum radius ζ^* such that $|\eta^{\zeta^*}(o_i)| \in \{3, 5, 7\}$ and calculate the mean, standard deviation, and disorder over all $o_i \in \mathbf{N}$. A total of 25 nuclear statistics ($\mathbf{f}_{\text{NS}}$) are extracted for each scene $\mathcal{C}$ (Table 5.1).

Type	Name
VD (13)	Total area of polygons
	Polygon area: mean, std dev, min/max ratio, disorder
	Polygon perimeter: mean, std dev, min/max ratio, disorder
	Polygon chord length: mean, std dev, min/max ratio, disorder
DT (8)	Triangle side length: mean, std dev, min/max ratio, disorder
	Triangle area: mean, std dev, min/max ratio, disorder
MST (4)	Edge length: mean, std dev, min/max ratio, disorder
NS (25)	Nuclear density
	Distance to 3 nearest nuclei: mean, std dev, disorder
	Distance to 5 nearest nuclei: mean, std dev, disorder
	Distance to 7 nearest nuclei: mean, std dev., disorder
	# nuclei in 10 µm radius: mean, std dev, disorder
	# nuclei in 20 µm radius: mean, std dev, disorder
	# nuclei in 30 µm radius: mean, std dev, disorder
	# nuclei in 40 µm radius: mean, std dev, disorder
	# nuclei in 50 µm radius: mean, std dev, disorder

Table 5.1: The 50 nuclear architecture features used in this paper, derived from Voronoi (VG), Delaunay triangulation (DT), and minimum spanning tree (MST) graphs, as well as nuclear statistics (NS).

5.3 Quantification of nuclear texture via Haralick co-occurrence features

Using the nuclear mask to restrict analysis to the desired region, Haralick co-occurrence features [32, 45] are extracted from each image. First, the image is transformed from the RGB color space to the HSI color space since the latter is more similar to the manner in which humans perceive color [101]. At each relevant pixel, a co-occurrence matrix is constructed to quantify the frequency of pixel intensities in a fixed neighborhood. A set of 13 Haralick features [32] are extracted from the co-occurrence matrices (Contrast Energy, Contrast Inverse Moment, Contrast Average, Contrast Variance, Contrast Entropy, Intensity Average, Intensity Variance, Intensity Entropy, Entropy, Energy, Correlation, and two Information Measures of Correlation), from which the mean, standard deviation, and disorder statistics are calculated for each image (see

Figure 7.2 for examples of nuclear texture responses). This task is repeated for each of the three channels in the HSI color space, resulting in a total of 117 nuclear texture features $\mathbf{f}_{\mathrm{NT}}(\mathcal{C})$ for each image scene $\mathcal{C}$.

5.4 Tubule density in breast cancer DP images

The ability to distinguish low and high tubule density (Figure 1.3) in H & E stained BCa histopathology is a key component of the mBR grading system and, hence, predicting patient outcome. Using the class predictions (i.e. tubule ω_1 or non-tubule ω_2) for individual lumen $\theta \in \{\omega_1, \omega_2\}$ calculated in Section 4.6.6, we are able to evaluate the degree of tubule formation across the entire image. We define tubule density for each image as the fraction of nuclei arranged in tubules

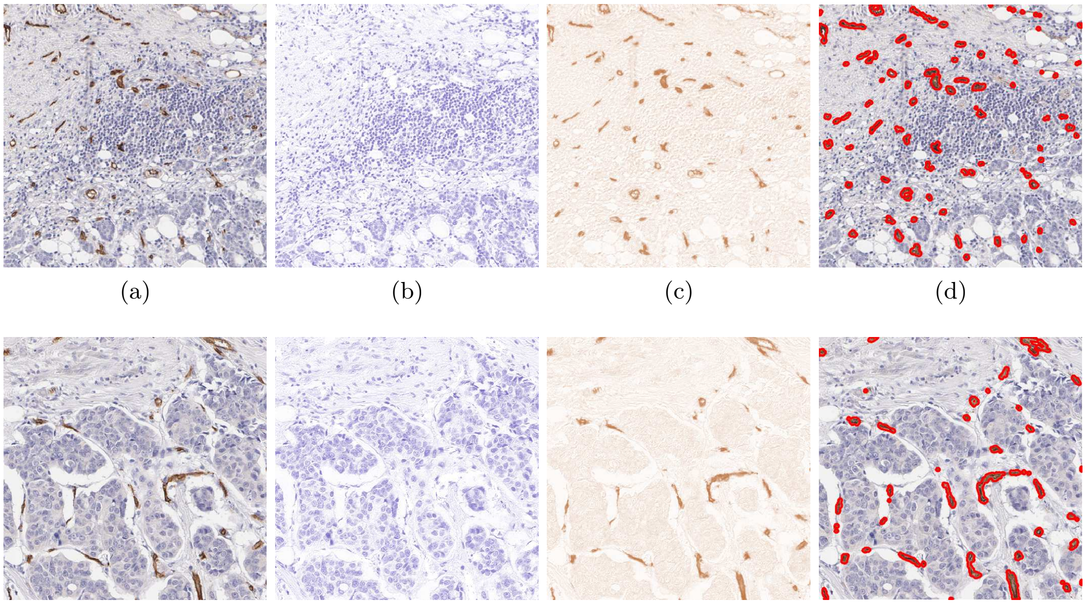
$$f_{\mathrm{TD}} = \frac{|\hat{\mathbf{N}}|}{|\mathbf{N}|}, \tag{5.1}$$

where $\hat{\mathbf{N}} = \{\mathbf{N}^{\ell} : o^{\ell} \in \mathbf{N},\ \theta(o^{\ell}) = \omega_1\}$ represents the set of all nuclei contained within the O'Callaghan neighborhoods of true lumen, $\mathbf{N}$ is the set of all nuclei in the image, and $|\cdot|$ denotes set cardinality.

5.5 Vascular density in CD34 IHC-stained DP

The CD34 protein is a popular indicator of angiogenesis and, hence, tumor growth and metastasis [102]. Previously, both qualitative [103] and quantitative [104] assessments of CD34 IHC stained slides have characterized IHC staining via "hotspots", i.e. manually selected FOVs; yet, the pitfalls associated with manual FOV selection (described in Section 2.5) suggest that hotspot-based predictions may not accurately represent CD34 expression in an entire slide. In this work, we quantify angiogenic activity by isolating the brown diaminobenzidine (DAB) compound signifying CD34 expression and use it to calculate the density of vascular formation via the following steps.

Step 1: Color deconvolution [90] is used to split the image into channels representing the DAB and hematoxylin stains (Figures 5.2(b), (c), (f), (g)).


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.2: (a), (e) CD34 IHC stained images are separated into (b), (f) hematoxylin and (c), (g) DAB channels via color deconvolution. The DAB channel is thresholded to isolate (d), (h) segmented regions expressing the CD34 protein.

Step 2: The DAB channel is thresholded to produce a set of brown pixels corresponding to angiogenic vessels (Figures 5.2(d), (h)).

Step 3: Vascular density (f_{VD}) is defined as fraction of brown pixels within region of cancer extent from an image.

5.6 Feature selection via Minimum Redundancy Maximum Relevance

We mitigate the limitations of large feature sets by employing the Minimum Redundancy Maximum Relevance (mRMR) feature selection scheme [34]. Given a feature set $\mathbf{f}$, the mRMR scheme identifies a subset $\bar{\mathbf{f}} \subset \mathbf{f}$ that maximizes "relevance" and minimizes "redundancy" between individual features. In practice, feature f_j is incrementally included in $\bar{\mathbf{f}}$ based on the criteria

$$f_j = \operatorname*{argmax}_{f_j \in \mathbf{f} - \bar{\mathbf{f}}} \left[I(f_j, \mathcal{Y}) - \frac{1}{|\bar{\mathbf{f}}| - 1} \sum_{f_i \in \bar{\mathbf{f}}} I(f_j, f_i) \right], \tag{5.2}$$

where I is mutual information, $\mathcal{Y}$ is the class label associated with a given sample, and $|\bar{\mathbf{f}}|$ represents the cardinality of selected feature set. In this work, relevant features are isolated from both nuclear architecture $\bar{\mathbf{f}}_{\mathrm{NA}} \subset \mathbf{f}_{\mathrm{NA}}$ and nuclear texture $\bar{\mathbf{f}}_{\mathrm{NT}} \subset \mathbf{f}_{\mathrm{NT}}$ feature sets based on their ability to distinguish BCa histopathology slides with low, intermediate, and high mBR grades.

Chapter 6

Multi-field-of-view sampling and classification framework

6.1 Specific notation used in this chapter

For all methods, an image scene $\mathcal{C} = (C, \mathbf{g})$ is defined as a 2D set of pixels $c \in C$ with associated vectorial function $\mathbf{g}$ assigning the RGB color space and class label $\mathcal{Y}(\mathcal{C}) \in \{0, 1\}$. For each $\mathcal{C}$ and FOV size $\tau \in T$, a grid containing FOVs $D^\tau = \{d_1^\tau, d_2^\tau, \ldots, d_{M(\tau)}^\tau\}$ is constructed, where $d_m^\tau \subset C, m \in \{1, 2, \ldots, M(\tau)\}$ is a square FOV with edge length of τ pixels and $M(\tau)$ is the total number of FOVs for a given τ. We define $\mathbf{f}(d_m^\tau)$ as the function that extracts features from each d_m^τ. Grid construction and feature extraction are repeated likewise for each $\tau \in T$.

6.2 Theory of multi-field-of-view classification

A consensus predictor over multiple FOV sizes is defined as $\mathbf{H}(\mathbf{D}; \mathbf{f}) = E_\tau \left[H(D^\tau; \tau, \mathbf{f}) \right]$, where $\mathbf{D} = \{D^\tau : \tau \in T\}$ is the collective data over all FOV sizes, $H(D^\tau; \tau, \mathbf{f})$ is a meta-prediction at FOV size $\tau \in T$, and E_τ is the expectation of $H(D^\tau, \tau; \mathbf{f})$ at FOV size $\tau \in T$. The mean squared error of classification at an individual FOV size is given by $e_\tau = E_\tau \left[\mathcal{Y} - H(D^\tau; \tau, \mathbf{f}) \right]^2$ and the error of the consensus predictor is given by $e_A = \left[\mathcal{Y} - \mathbf{H}(\mathbf{D}; \mathbf{f}) \right]^2$.

Proposition 1. *Given independent classifiers at FOV sizes $\tau \in T$, $e_\tau \geq e_A$.*

Proof.

$$\begin{aligned} e_\tau &= E_\tau \left[\mathcal{Y} - H(D^\tau; \tau, \mathbf{f}) \right]^2 \\ &= \mathcal{Y}^2 - 2\mathcal{Y} E_\tau \left[H(D^\tau; \tau, \mathbf{f}) \right] + E_\tau \left[H^2(D^\tau; \tau, \mathbf{f}) \right] \end{aligned}$$

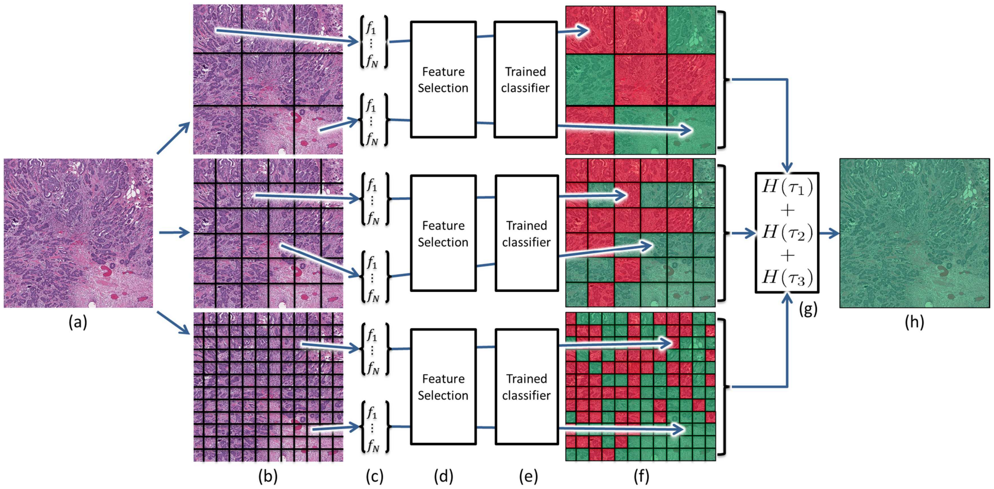


Figure 6.1: A flowchart outlining the methodological steps of the multi-FOV classifier in terms of its application to differentiating mBR grade in ER+ BCa histopathology. First, (a) a histopathology slide is first divided into (b) FOVs of various sizes. (c) Image features that quantify mBR grade phenotype are extracted from each FOV and (d) a feature selection scheme is used to identify salient features at each FOV size. (e) Pre-trained classifiers are used to predict (f) mBR grade for each FOV (illustrated by red and green squares). (g) Predictions for individual FOVs are aggregated to achieve a class prediction $H(\tau)$ for an entire FOV size τ. (h) Class predictions from FOV sizes are combined to achieve a final classification result for the entire ER+ BCa histopathology slide.

$$
\begin{aligned}
\text{Since} \quad & E_\tau\left[H^2(D^\tau;\tau,\mathbf{f})\right] \geq \left[E_\tau\left[H(D^\tau;\tau,\mathbf{f})\right]\right]^2, \\
& \geq \mathcal{Y}^2 - 2\mathcal{Y}E_\tau\left[H(D^\tau;\tau,\mathbf{f})\right] + \left[E_\tau\left[H(D^\tau;\tau,\mathbf{f})\right]\right]^2 \\
& \geq \mathcal{Y}^2 - 2\mathcal{Y}\mathbf{H}(\mathbf{D};\mathbf{f}) + \mathbf{H}^2(\mathbf{D};\mathbf{f}) \\
& \geq \left[\mathcal{Y} - \mathbf{H}(\mathbf{D};\mathbf{f})\right]^2 \\
& \geq e_A
\end{aligned}
$$

□

Note that the consensus classifier for multiple FOV sizes is similar to Bagging [105]. In this approach, independent predictors at different FOV sizes are used as the "weak" learners and combined to build the "strong" consensus result. To this end, Proposition 1 ensures that the consensus error e_A will always be less than the mean error e_τ of individual FOV size classifiers.

Algorithm 3 MultiFOV()

Input: Image $\mathcal{C}$. FOV sizes $T = \{t_1, t_2, \ldots, t_N\}$. Classifier $\mathbf{h}(d_m^\tau; \tau, \mathbf{f})$ for each $\tau \in T$.
Output: Multi-FOV classification $\mathbf{H}(\mathbf{D}; \mathbf{f})$ for image $\mathcal{C}$.

```
for all $\tau \in T$ do
    From $\mathcal{C}$, define $M(\tau)$ FOVs $D^\tau = \{d_1^\tau, d_2^\tau, \ldots, d_{M(\tau)}^\tau\}$.
    Extract features $\mathbf{f}$ from $d_m^\tau$, $\forall m \in \{1, 2, \ldots, M(\tau)\}$.
    Initial classification $\mathbf{h}(d_m^\tau; \tau, \mathbf{f})$ of each $d_m^\tau$.
    For all FOVs $D^\tau$ at size $\tau$, make class prediction $H(D^\tau; \tau, \mathbf{f}) = \frac{1}{M(\tau)} \sum_{m=1}^{M(\tau)} \mathbf{h}(d_m^\tau; \tau, \mathbf{f})$.
end for
Across all FOV sizes $\tau \in T$, make multi-FOV prediction $\mathbf{H}(\mathbf{D}; \mathbf{f}) = \frac{1}{N} \sum_{\tau \in T} H(D^\tau; \tau, \mathbf{f})$.
```

6.3 Implementation of multi-FOV classifier for whole-slide DP images

The multi-FOV framework (Figure 6.1) is designed to classify large, heterogeneous images in an automated and unbiased fashion as described in Algorithm 3 [2,3,7]. For a single slide $\mathcal{C}$, a pre-trained classifier $\mathbf{h}(d_m^\tau; \tau, \mathbf{f}) \in \{0, 1\}$ is first used to assign an initial class prediction for each individual FOV d_m^τ with associated features $\mathbf{f}$. Predictions are aggregated (i.e. mean prediction) for all FOVs D^τ at a single size $\tau \in T$ to achieve a combined prediction $H(D^\tau; \tau, \mathbf{f})$. Subsequently, the multi-FOV classification $\mathbf{H}(\mathbf{D}; \mathbf{f})$, where $\mathbf{D} = \{D^\tau : \forall \tau \in T\}$ is the collective data over all FOV sizes, is achieved via a consensus prediction across all FOV sizes. In this work, consensus is achieved via averaging of $H(D^\tau; \tau, \mathbf{f}), \forall \tau \in T$.

6.4 Multi-parametric extension for additional channels of histopathological data

The multi-FOV framework is readily extensible to additional types of DP images, allowing for the integration of prognostic information from a variety of complementary histolgical sources [2]. In this work, we show how two channels of histopathological data from the same tumor (e.g. H & E stained histology and IHC-stained histology) can be combined via the following steps.

Step 1: Perform *MultiFOV()* (Algorithm 3) for the first data channel and save resulting class decision $\mathbf{H}_1 \in \{0, 1\}$.

Step 2: Perform *MultiFOV()* (Algorithm 3) for the second data channel and save resulting class decision $\mathbf{H}_2 \in \{0, 1\}$.

Step 3: Generate a decision-level prediction $\hat{\mathbf{H}} = \mathbf{H}_1 \wedge \mathbf{H}_2 \in \{0, 1\}$ based on the independent class predictions. Note that the $\wedge$ operation is defined as "logical AND", whereby $\hat{\mathbf{H}} = 1$ if both $\mathbf{H}_1 = 1$ and $\mathbf{H}_2 = 1$. Conversely, $\hat{\mathbf{H}} = 0$ if either $\mathbf{H}_1 = 0$ or $\mathbf{H}_2 = 0$.

6.5 Experimental validation of multi-FOV framework

The theory set forth in Proposition 1, which suggests that the error rate of the multi-FOV classifier will always be less than a majority of the error rates from its constituent FOV sizes, is evaluated in the context of distinguishing mBR grade from DP images of ER+ breast cancers.

6.5.1 Data cohort of ER+ breast cancers

Anonymized BCa histopathology slides were obtained from 126 patients (46 low mBR, 60 intermediate mBR, 20 high mBR) at the Hospital of the University of Pennsylvania (Philadelphia, PA) and The Cancer Institute of New Jersey (New Brunswick, NJ). All slides were digitized via a whole slide scanner at 10x magnification (1 µm/pixel resolution). Each slide is accompanied by (1) an annotation corresponding to regions containing IDC and (2) mBR grade as determined by an expert pathologist. Note that commonly accepted clinical cutoffs are used to define the low (mBR 3-5), intermediate (mBR 6-7), and high (mBR 8-9) grade classes used as ground truth in this work. The multi-FOV framework is evaluated via a series classification tasks to distinguish DP slides with low vs. high mBR grade, low vs. intermediate mBR grade, and intermediate vs. high mBR grade. In addition, a wide range of FOV sizes $T = \{4000, 2000, 1000, 500, 250\}$µm was selected to capture different aspects of tissue morphology [2, 3, 7].

FOV size (τ)	low vs. high	low vs. intermed.	intermed. vs. high
250	0.79 ± 0.054	0.65 ± 0.043	0.51 ± 0.029
500	0.79 ± 0.041	0.67 ± 0.041	0.56 ± 0.053
1000	0.82 ± 0.027	0.69 ± 0.051	0.61 ± 0.044
2000	0.80 ± 0.057	0.69 ± 0.023	0.57 ± 0.073
4000	0.73 ± 0.072	0.62 ± 0.044	0.58 ± 0.011
multi-FOV	**0.88 ± 0.028**	**0.73 ± 0.032**	**0.74 ± 0.04**

Table 6.1: Mean and standard deviation classification accuracies (over 20 trials of 3-fold cross-validation) are reported for individual FOV sizes and the multi-FOV classifier using features describing nuclear architecture.

6.5.2 Experimental design and results

The multi-FOV framework is applied as illustrated in Figure 6.1 using 50 image features that characterize nuclear architecture ($\mathbf{f}_{\text{NA}}$ described in Section 5.2). In order to avoid issues associated with a high-dimensional feature space, 5 features are selected independently at each FOV size via the mRMR algorithm (Section 5.6). Classification is performed using the random forest classifier [99] in conjunction with randomized 3-fold cross-validation. The cross-validation scheme is used to mitigate bias in the selection of training and testing samples by randomly dividing the dataset into 3 partitions. Data from two partitions are used for feature selection and classifier training, while the partition is used for evaluation. This process is repeated 3 times so that all samples are evaluated exactly once. In this study, the mean and standard deviation of classification accuracy for individual FOV sizes (H) and the multi-FOV result ($\mathbf{H}$) are reported over 20 trials of cross-validation (Table 6.1).

6.5.3 Discussion

The classification results shown in Table 6.1 clearly demonstrate the ability of the multi-FOV approach to outperform the majority of individual FOV sizes. In fact, application of the Wilcoxon rank-sum test [88] shows that the multi-FOV classifier performs significantly better ($p < 0.0001$) than all individual FOV sizes for distingushing both patients with low vs. high and intermediate vs. high mBR grades. Similarly, the multi-FOV classifier for distinguishing patients with low vs. intermediate mBR grade performed significantly better than all FOV sizes except for $\tau = 1000$, which yielded a p-value of 0.0745. Note that all p-values have been corrected for multiple comparisons using the conservative Bonferroni method [89]. Hence, these results serve as experimental validation of the theoretical concepts behind the multi-FOV framework presented in Section 6.2.

Chapter 7

Predicting and comparing large-scale classifier performance with limited training data

7.1 Commonly used notation in this chapter

For all experiments, a dataset $\mathcal{D}$ is divided into independent training $\mathcal{N} \subset \mathcal{D}$ and a testing $\mathcal{T} \subset \mathcal{D}$ pools, where $\mathcal{N} \cap \mathcal{T} = \emptyset$. Note that the datasets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ defined here are used only in this chapter and are different from the datasets employed in Chapters 8 and 9. The class label of a sample $x \in \mathcal{D}$ is denoted by $\mathcal{Y}_t \in \{\omega_1, \omega_2\}$. A set of training set sizes $\mathbf{N} = \{n_1, n_2, \ldots, n_N\}$, where $1 \leq n \leq |\mathcal{N}|$ and $| \cdot |$ denotes set cardinality. Additional parameters for the extended RRS approach presented in this chapter include T_1 subsets for the subsampling test, T_2 randomized subsets for the permutation test, and K folds for cross-validation sampling. Other commonly used notation can be found in Appendix A.

7.2 Subsampling Test to Calculate Error Rates for Multiple Training Set Sizes

The estimation of classifier performance first requires the construction of multiple classifiers trained on repeated subsampling of the limited dataset. For each training set size $n \in \mathbf{N}$, a total of T_1 subsets $\mathcal{S} \in \mathbb{R}^{n \times T1}$ are created by randomly sampling the training pool $\mathcal{N}$. For each $n \in \mathbf{N}$ and $i \in \{1, 2, \ldots, T_1\}$, the subset $S_i(n) \in \mathcal{S}$ is used to train a corresponding classifier $H_i(n)$. Each $H_i(n)$ is evaluated on the entire testing set $\mathcal{T}$ to produce an error rate $e_i(n)$. The mean error rate for each $n \in \mathbf{N}$ is calculated as

$$\bar{e}(n) = \frac{1}{T_1} \sum_{i}^{T_1} e_i(n). \tag{7.1}$$

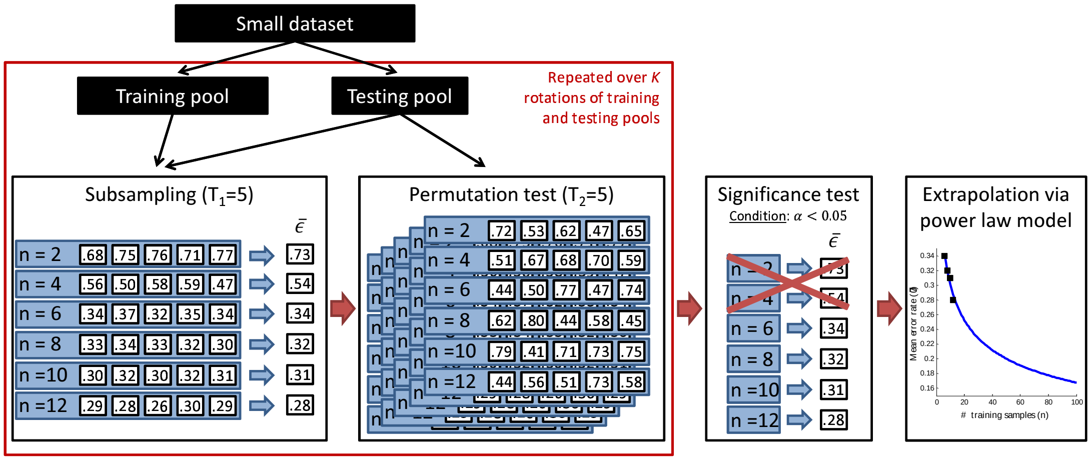

Small dataset
Training pool
Testing pool
Repeated over K rotations of training and testing pools
Subsampling (T1=5)
ϵ̄
n = 2 .68 .75 .76 .71 .77 .73
n = 4 .56 .50 .58 .59 .47 .54
n = 6 .34 .37 .32 .35 .34 .34
n = 8 .33 .34 .33 .32 .30 .32
n =10 .30 .32 .30 .32 .31 .31
n =12 .29 .28 .26 .30 .29 .28
Permutation test (T2=5)
n = 2 .72 .53 .62 .47 .65
n = 4 .51 .67 .68 .70 .59
n = 6 .44 .50 .77 .47 .74
n = 8 .62 .80 .44 .58 .45
n =10 .79 .41 .71 .73 .75
n =12 .44 .56 .51 .73 .58
Significance test
Condition: α < 0.05
ϵ̄
n = 2 .73
n = 4 .54
n = 6 .34
n = 8 .32
n =10 .31
n =12 .28
Extrapolation via power law model
Mean error rate (ϵ̄)
0.34
0.32
0.3
0.28
0.26
0.24
0.22
0.2
0.18
0.16
0
20
40
60
80
100
# training samples (n)

Figure 7.1: A flowchart describing the methodology used in this chapter. First, a dataset is partitioned into training and testing pools using a K-fold sampling strategy. Each of the K training pools undergoes repeated random sampling (RRS), in which error rates are calculated at different training set sizes via a subsampling procedure. A permutation test is used to identify statistically significant error rates, which are then used to extrapolate learning curves and predict error rates for larger datasets.

7.3 Permutation Test to Evaluate Statistical Significance of Error Rates

To ensure the statistical significance of the mean error rates $\bar{e}(n)$ calculated in Equation 7.1, the performance of training set $S_i(n)$ is compared against the performance of randomly labeled training data. For each $S_i(n) \in \mathcal{S}$, a total of T_2 random training sets $\hat{\mathcal{S}} \in \mathbb{R}^{n \times T_1 \times T_2}$ are created in which each sample is assigned a randomized class label $\mathcal{Y}_r \in \{\omega_1, \omega_2\}$. For each $n \in \mathbf{N}$, $i \in \{1, 2, \ldots, T_1\}$, and $j \in \{1, 2, \ldots, T_2\}$, the subset $\hat{S}_{i,j}(n) \in \hat{\mathcal{S}}$ is used to train a corresponding classifier $\hat{H}_{i,j}(n)$. Each $\hat{H}_{i,j}(n)$ is evaluated on the entire testing set $\mathcal{T}$ to produce an error rate $\hat{e}_{i,j}(n)$. For each n, a p-value

$$P_n = \frac{1}{T_1} \frac{1}{T_2} \sum_{i=1}^{T_1} \sum_{j=1}^{T_2} \theta(\bar{e}(n) - \hat{e}_{i,j}(n)), \tag{7.2}$$

where $\theta(z) = 1$ if $z \geq 0$ and 0 otherwise. P_n is calculated as the fraction of randomly-labeled classifiers $\hat{H}_{i,j}(n)$ with error rates $\bar{e}_{i,j}(n)$ exceeding the mean error rate $\bar{e}(n), \forall n \in \mathbf{N}$. The mean error rate $\bar{e}(n)$ is deemed to be valid for model-fitting only if $P_n < 0.05$, i.e. there is a statistically significant difference between $\bar{e}(n)$ and $\{\hat{e}_{i,j}(n), \forall i \in \{1, 2, \ldots, T_1\}, \forall j \in \{1, 2, \ldots, T_2\}\}$. Hence, the set of valid training set sizes $\mathbf{M} = \{n :$

$n \in \mathbf{N}, P_n < 0.05\}$ includes only those $n \in \mathbf{N}$ that have passed the significance test.

7.4 Cross-Validation Strategy for Selection of Training and Testing Pools

The selection of training $\mathcal{N}$ and testing $\mathcal{T}$ pools from the limited dataset $\mathcal{D}$ is governed by a K-fold cross-validation strategy. In this work, the dataset $\mathcal{D}$ is partitioned into $K = 4$ pools in which one pool is used for evaluation while the remaining $K - 1$ pools are used for training to produce mean error rates $\bar{e}_k(n)$, where $k \in \{1, 2, \ldots, K\}$. The pools are then rotated and the subsampling and permutation tests are repeated until all pools have been evaluated exactly once. This process is repeated over R cross-validation trials, yielding mean error rates $\bar{e}_{k,r}(n)$ where $r \in \{1, 2, \ldots, R\}$. For all training set sizes that have passed the significance test, i.e. $\forall n \in \mathbf{M}$, power law curves are generated from a comprehensive mean error rate

$$\bar{\mathbf{e}}(n) = \frac{1}{K}\frac{1}{R}\sum_{k=1}^{K}\sum_{r=1}^{R}\bar{e}_{k,r}(n), \tag{7.3}$$

calculated over all cross-validation folds $k \in \{1, 2, \ldots, K\}$ and iterations $r \in \{1, 2, \ldots, R\}$.

7.5 Estimation of Power Law Model Parameters

The power law model [82] describes the relationship between error rate and training set size

$$\bar{\mathbf{e}}(n) = an^{-\alpha} + b, \tag{7.4}$$

where $\bar{\mathbf{e}}(n)$ is the comprehensive mean error rate (Equation 7.3) for training set size n, a is the learning rate, and α is the decay rate. The Bayes error rate b is defined as the lowest possible error given an infinite amount of training data [41]. The model parameters a, α, and b are calculated by solving the constrained non-linear minimization problem

$$\min_{a,\alpha,b}\sum_{m=1}^{|\mathbf{M}|}(an_m S^{-\alpha} + b - \bar{\mathbf{e}}(n))^2, \tag{7.5}$$

where $a, \alpha, b \geq 0$.

7.6 Extension of Error Rate Prediction to Pixel- and Voxel-level Data

Application of the model presented in this work to patient-level medical imaging data, where each patient is described by a single set of features, is relatively well-understood. Yet disease classification in radiological data (e.g. MRI) occurs at the pixel-level, in which each patient has pixels from both classes (e.g. diseased and non-diseased states) and each pixel is characterized by a set of features. The methodology presented in this work can be extended to such pixel- or voxel-level data by first selecting training set sizes $\mathbf{N}$ at the patient-level. Definition of the K training and testing pools as well as creation of each subsampled training set $S_i(n) \in \mathcal{S}$ are also performed at the patient-level. Training of the corresponding classifier $H_i(n)$, however, is performed at the pixel-level by aggregating pixels for all patients in $S_i(n)$. A similar aggregation is done for all patients in the testing pool $\mathcal{T}$. By ensuring that all pixels from a given patient remain together, we are able to perform extrapolation of pixel-wise data while avoiding the classification bias that occurs when pixels from a single patient span both training and testing sets.

7.7 Experimental Design

Our methodology is evaluated on 3 classification tasks traditionally affected by limitations in the availability of imaging data (Table 7.1). All experiments have a number of parameters in common, including $T_1 = 50$ subsampling trials, $T_2 = 50$ permutation trials, $K = 4$ cross-validation folds, and $R = 10$ cross-validation trials. In addition, all experiments employ the k-nearest neighbor (kNN), naive Bayes (NB), and Support Vector Machine (SVM) classifiers. A more detailed description of each classifier is presented in Appendix B. In each experiment, validation is performed via leave-one-out (LOO) classification on a larger dataset, which allows us to maximize the number of training samples used for classification while yielding the expected lower bound of the error rate.

Notation	Description	Samples (train / valid.)
$\mathcal{D}_1$	Prostate: Cancer detection on histopathology	100 / 500
$\mathcal{D}_2$	Breast: Cancer grading on histopathology	46 / 116
$\mathcal{D}_3$	Prostate: Cancer detection on MRS	16 / 34

Table 7.1: List of the breast cancer and prostate cancer datasets used in this study. For $\mathcal{D}_3$, training and testing sets are selected at the patient-level, while classification is performed at the metavoxel-level by using all metavoxels from both classes for a specified patient.

7.7.1 Experiment 1: Identifying Cancerous Tissue in Prostate Cancer Histopathology

Automated systems for detecting prostate cancer on biopsy specimens have the potential to act as (1) a triage mechanism to help pathologists spend less time analyzing samples without cancer and (2) an initial step for decision support systems that aim to quantify disease aggressiveness via automated Gleason grading [38]. Dataset $\mathcal{D}_1$ comprises hematoxylin and eosin (H & E) stained needle-core biopsies of prostate tissue digitized at 20x optical magnification on a whole-slide digital scanner. Regions corresponding to prostate cancer were manually delineated by a pathologist and used as ground truth. Slides were divided into non-overlapping 30 × 30-pixel tissue regions and converted to a grayscale representation. A total of 927 features including first-order statistical, Haralick co-occurrence [32], and steerable Gabor filter features were extracted from each image [45] (Table 7.2). Due to the small number of training samples used in this study, the feature set was first reduced to two descriptors via the minimum redundancy maximum relevance (mRMR) feature selection scheme [34], primarily to avoid the curse of dimensionality [41]. A relatively small dataset of 100 image regions, with training set sizes $\mathbf{N} = \{25, 30, 35, 40, 45, 50, 55\}$, was used to extrapolate error rates (Table 7.1). LOO cross-validation was subsequently performed on a larger dataset comprising 500 image regions.

Features	Parameters
Texture: Gray-level (Average, Median, Standard Deviation, Range, Sobel, Kirsch, Gradient, Derivative)	window sizes: $\{3, 5, 7\}$
Texture: Haralick co-occurrence (Joint Entropy, Energy, Inertia, Inverse Difference Moment, Correlation, Measurements of Correlation, Sum Average, Sum Variance, Sum Entropy, Difference Average, Difference Variance, Difference Entropy, Shade, Prominence, Variance)	window sizes: $\{3, 5, 7\}$
Texture: steerable Gabor filter responses (cosine and sine components combined)	window sizes: $\{3, 5, 7\}$ frequency shift: $\{0, 1, \ldots, 7\}$ orientations: $\{0, \frac{\pi}{8}, \frac{2\pi}{8}, \ldots, \frac{7\pi}{8}\}$

Table 7.2: A summary of all features extracted from prostate cancer histopathology images in dataset $\mathcal{D}_1$. All textural features were extracted separately for red, green, and blue color channels.

7.7.2 Experiment 2: Distinguishing High and Low Tumor Grade in Breast Cancer Histopathology

Nottingham, or modified Bloom-Richardson (mBR), grade is routinely used to characterized tumor differentiation in breast cancer (BCa) histopathology [18]; yet, it is known to suffer from high inter- and intra-pathologist variability [23]. Hence, researchers have aimed to develop quantitative and reproducible classification systems for differentiating mBR grade in BCa histopathology [3]. Dataset $\mathcal{D}_2$ comprises 2000×2000 image regions taken from H & E stained histopathology specimens of breast tissue digitized at 20x optical magnification on a whole-slide digital scanner. Ground truth for each image was determined by an expert pathologist to be either low (mBR < 6) or high (mBR > 7) grade. First, boundaries of 30-40 representative epithelial nuclei were manually segmented in each image region (Figure 7.2). Using the segmented boundaries, a total of 2343 features were extracted from each nucleus to quantify both nuclear morphology and nuclear texture (Table 7.3). A single feature vector was subsequently defined for each image region by calculating the median feature values of all constituent nuclei. Similar to Experiment 1, mRMR feature selection was used to isolate the two

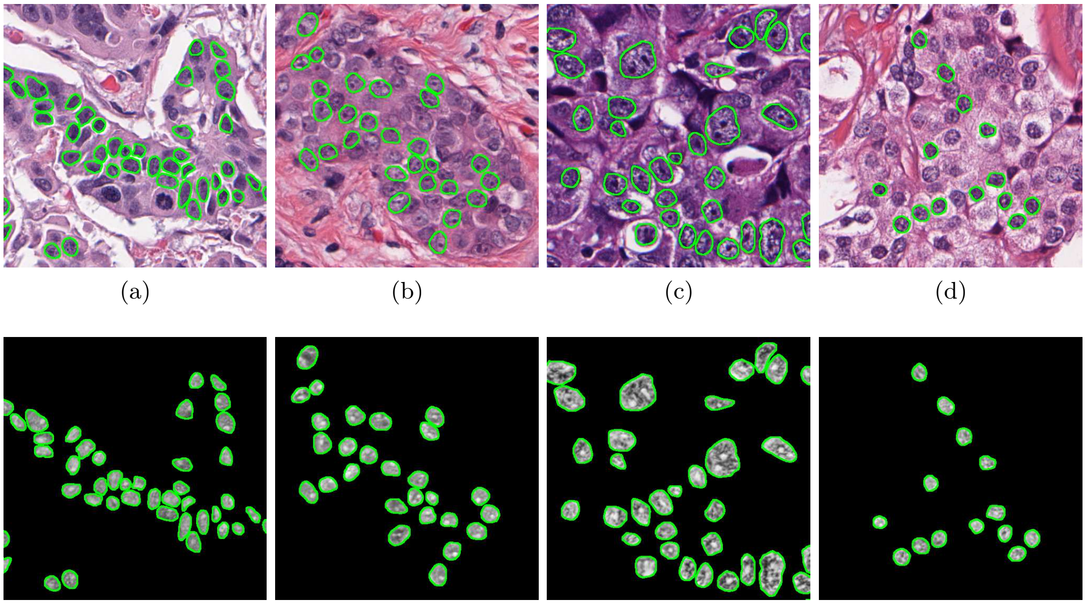


Figure 7.2: Examples of (a), (b) low mBR grade and (c), (d) high mBR grade BCa histopathology images from dataset $\mathcal{D}_2$ shown with boundary annotations (green outline) for exemplar nuclei. A variety of morphological and textural features are extracted from the nuclear regions, including (e)-(h) the Sum Variance Haralick textural response.

most important descriptors. Error rates were extrapolated from a small dataset comprising 45 images with training set sizes $\mathbf{N} = \{20, 22, 24, 26, 28, 30, 32\}$, while LOO cross-validation was subsequently performed on a larger dataset comprising 116 image regions (Table 7.1).

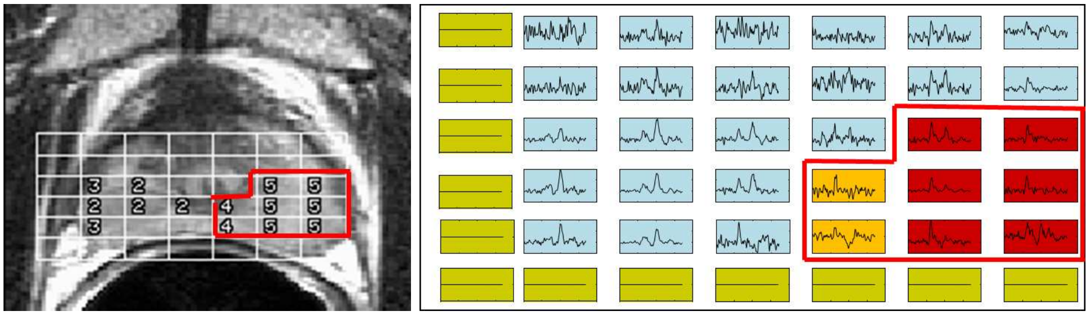
7.7.3 Experiment 3: Identifying Cancerous Metavoxels in Prostate Cancer Magnetic Resonance Spectroscopy

Magnetic resonance spectroscopy (MRS), a metabolic non-imaging modality that obtains the metabolic concentrations of specific molecular markers and biochemicals in the prostate, has previously been shown to supplement MRI in the detection of prostate cancer [106, 107]. These include choline, creatine, and citrate, and changes in their relative concentrations (choline/citrate or [choline+creatine)/citrate]), which have been shown to be linked to presence of prostate cancer [108]. Radiologists typically assess presence of prostate cancer on MRS by comparing ratios between choline, creatine, and citrate

Features	**Parameters**
Morphological (Area, Major Axis Length, Minor Axis Length, Eccentricity, Convex Area, Filled Area, Equivalent Diameter, Solidity, Extent, Perimeter, Area Overlap, Average Radial Ratio, Compactness, Convexity, Smoothness, Std. Dev. of Distance Ratio, Fourier Descriptors (6 rotations))	–
Texture: Gray-level (Average, Median, Standard Deviation, Range, Sobel, Kirsch, Gradient, Derivative)	window sizes: $\{3, 5, 7\}$
Texture: Local binary patterns	window size: 3 offsets: $\{0, 1, \ldots, 7\}$ directions: clockwise, counter-clockwise
Texture: Laws (pairwise convolution of Level, Edge, Spot, Wave, Ripple filters)	–
Texture: steerable Gabor filter responses (cosine and sine components are separate features)	window sizes: $\{3, 5, 9\}$ orientations: $\{0, \frac{\pi}{12}, \frac{2\pi}{12}, \ldots, \frac{6\pi}{12}\}$

Table 7.3: A summary of all features extracted from breast cancer histopathology images in dataset $\mathcal{D}_2$. All textural features were extracted separately for red, green, and blue color channels from the RGB color space and the hue, saturation, and intensity color channels from the HSV color space.

peaks to predefined normal ranges. Dataset $\mathcal{D}_3$ comprises 34 1.5 Tesla T2-weighted MRI and MRS studies obtained prior to radical prostatectomy, where the ground truth was defined (as cancer and benign metavoxels) via visual inspection of MRI and MRS by an expert radiologist [107] (Figure 7.3). Six MRS features were defined for each metavoxel by calculating expression levels for each metabolite as well as ratios between each pair of metabolites. Similar to Experiment 1, mRMR feature selection was used to identify the two most important features in the dataset. Error rates were extrapolated from a dataset of 16 patients using training set sizes $\mathbf{N} = \{2, 4, 6, 8, 10, 12\}$, followed by LOO cross-validation on a larger dataset of 34 patients (Table 7.1).



(a) (b)

Figure 7.3: (a) A study from dataset $\mathcal{D}_3$ showing an MR image of the prostate with MRS metavoxel locations overlaid. (b) For ground truth, each MRS spectrum is labeled as either cancerous (red and orange boxes) or benign (blue boxes). Green boxes correspond to metavoxels outside the prostate for which MRS spectra were suppressed during acquisition.

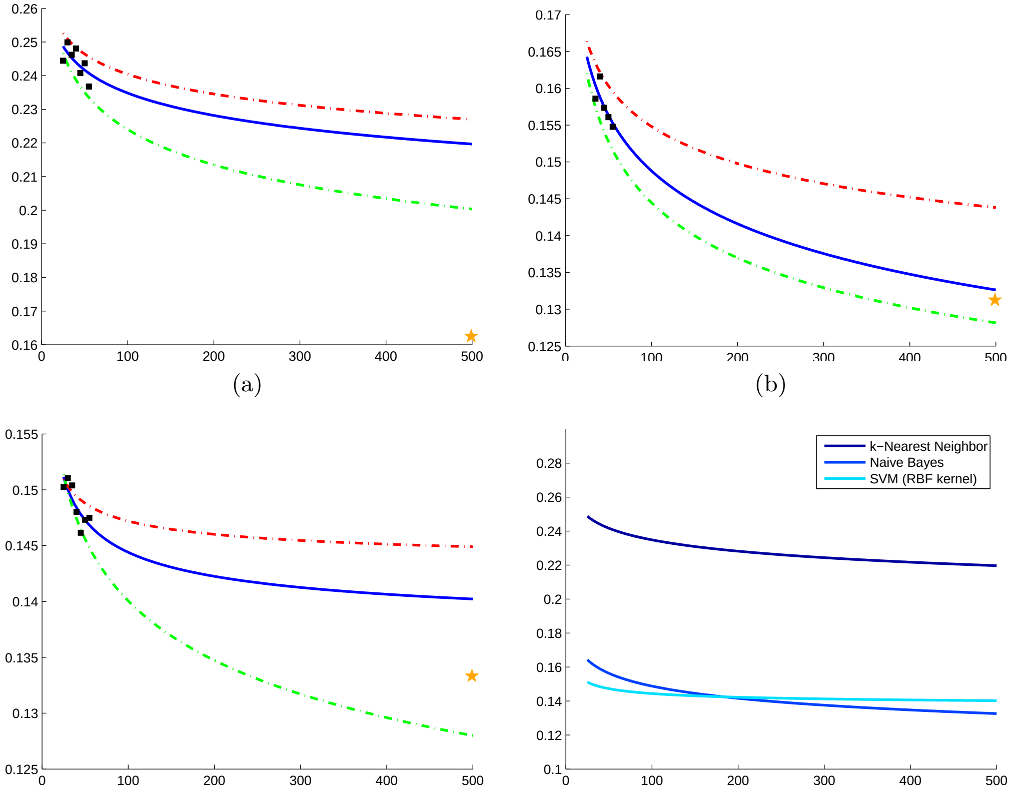
7.7.4 Comparison with Traditional RRS via Interquartile Range

This experiment employs dataset $\mathcal{D}_1$ and experimental parameters used in Experiment 1 within the traditional RRS approach. However, since traditional RRS does not use cross-validation, a total of $\hat{T}_1 = T_1 \cdot K \cdot R$ subsampling procedures are used to ensure that same number of classification tasks are performed for both approaches. Evaluation is performed via (1) comparison of the learning curves between the two methods and (2) the interquartile range (IQR), a measure of statistical variability defined as the difference between the 25th and 75th percentile error rates from the subsampling procedure.

7.8 Results and Discussion

7.8.1 Experiment 1: Distinguishing Cancerous and Non-Cancerous Regions in Prostate Histopathology

Error rates predicted by NB and SVM classifiers are similar to those from their LOO error rates of 0.1312 and 0.1333 (Figures 7.4(b), (c)). In comparison to the learning curves, the slightly lower error rate produced by the validation set is to be expected since the LOO classification is known to produce an overly optimistic estimate of the true error rate [109]. The kNN classifier appears to overestimate error considerably compared



(a) (b)

(c) (d)

Figure 7.4: Learning curves (blue line) generated for dataset $\mathcal{D}_1$ using mean error rates (black squares) calculated from (a) kNN, (b) NB, and (c) SVM classifiers. Each classifier is accompanied by curves for the 25th (green dashed line) and 75th (red dashed line) percentile of the error as well as LOO error on the validation cohort (yellow star). (d) A direct classifier comparison is made in terms of the mean error rate predicted by each learning curve in (a)-(c).

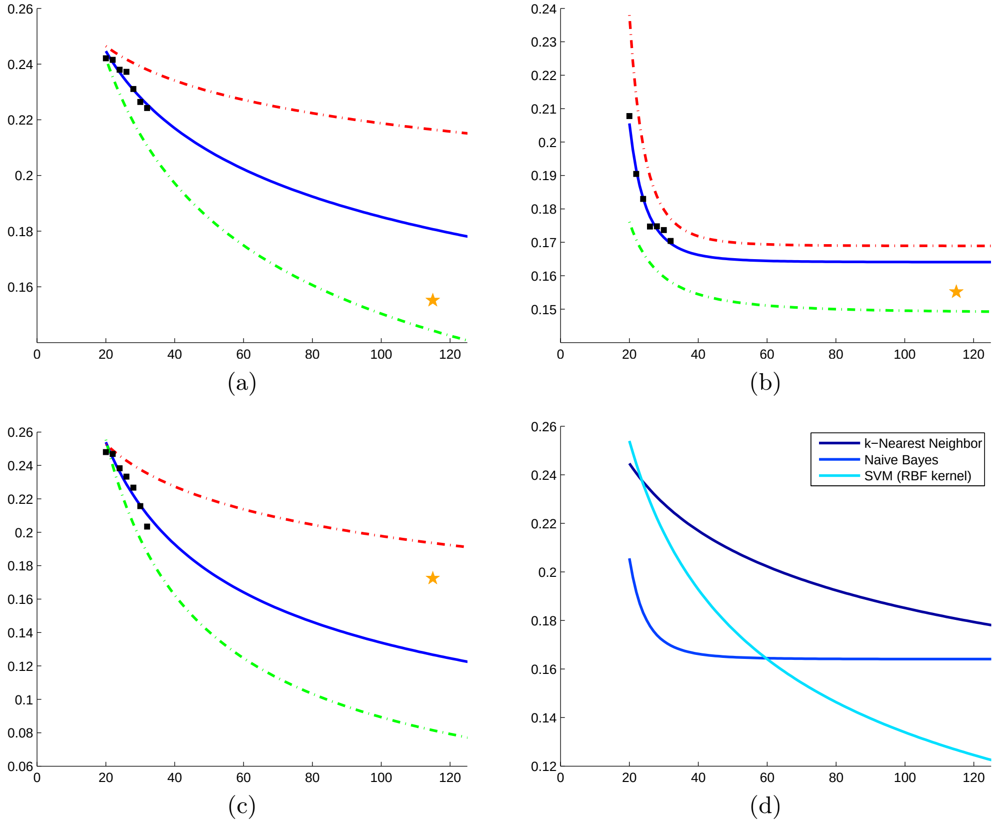
to the LOO error of 0.1560, which is not surprising because kNN is a non-parametric classifier that is expected to be more unstable for heterogeneous datasets (Figure 7.4(a)). Comparison across classifiers suggests that both NB and SVM will outperform kNN as dataset size increases (Figure 7.4(d)). Although the differences between the mean NB and SVM learning curves are minimal, the 25th and 75th percentile curves suggest that the prediction made by NB is more stable and has lower variance than the SVM prediction.

7.8.2 Experiment 2: Distinguishing Low and High Grade Cancer in Breast Histopathology

Learning curves from kNN and NB classifiers yield predicted error rates similar to their LOO cross-validation errors (0.1552 for both classifiers) as shown in Figures 7.5(a), (b). By contrast, while error rates predicted by the SVM classifier are reasonable (Figure 7.5(c)), they appear to underestimate the LOO error of 0.1724. One reason for this discrepancy may be the class imbalance present in the validation dataset (79 low grade and 37 high grade), since SVM classifiers have been demonstrated to perform poorly on datasets where the positive class (i.e. high grade) is underrepresented [110]. Similar to $\mathcal{D}_1$, a comparison between the learning curves reflects the superiority of both NB and SVM classifiers over the kNN classifier as dataset size increases (Figure 7.5(d)). However, the relationship between the NB and SVM classifiers is more complex. For small training sets, the NB classifier appears to outperform the SVM classifier; yet, the SVM classifier is predicted to yield lower error rates for larger datasets ($n > 60$). This suggests that the classifier yielding the best results for the smaller dataset may not necessarily be the optimal classifier as the dataset increases in size.

7.8.3 Experiment 3: Distinguishing Cancerous and Non-Cancerous Metavoxels in Prostate MRS

Similar to dataset $\mathcal{D}_1$, the LOO error for both the NB and SVM classifiers (0.2248 and 0.2468, respectively) fall within the range of the predicted error rates (Figures 7.6(b), (c)). Once again, the kNN classifier overestimates the LOO error (0.2628), which is most likely due to the high level of variability in the mean error rates used for extrapolation (Figure 7.6(a)). While both NB and SVM classifiers outperform the kNN classifier, their learning curves show a clearer separation between the extrapolated error rates for all dataset sizes, suggesting that the optimal classifier selected from the smaller dataset will hold true as even as dataset size increases (Figure 7.6(d)).



(a)

(b)

(c)

(d)

Figure 7.5: Learning curves (blue line) generated for dataset $\mathcal{D}_2$ using mean error rates (black squares) calculated from (a) kNN, (b) NB, and (c) SVM classifiers. Each classifier is accompanied by curves for the 25th (green dashed line) and 75th (red dashed line) percentile of the error as well as LOO error on the validation cohort (yellow star). (d) A direct classifier comparison is made in terms of the mean error rate predicted by each learning curve in (a)-(c).

7.8.4 Comparison with Traditional RRS

The quantitative results in Table 7.4 suggest that employing a cross-validation sampling strategy yields more consistent error rates. Traditional RRS yielded a mean IQR ($\overline{\text{IQR}}$) of 0.0297 across all $n \in \mathbf{N}$; whereas our approach demonstrated a lower $\overline{\text{IQR}}$ of 0.0070. Furthermore, a closer look at the learning curves for these error rates (Figure 7.7) suggests that traditional RRS is sometimes unable to accurately extrapolate learning curves. This phenomenon is most likely due to the high level of heterogeneity in medical imaging data and demonstrates the importance of rotating the training and testing pools to avoid biased error rates that do not generalize to larger datasets.

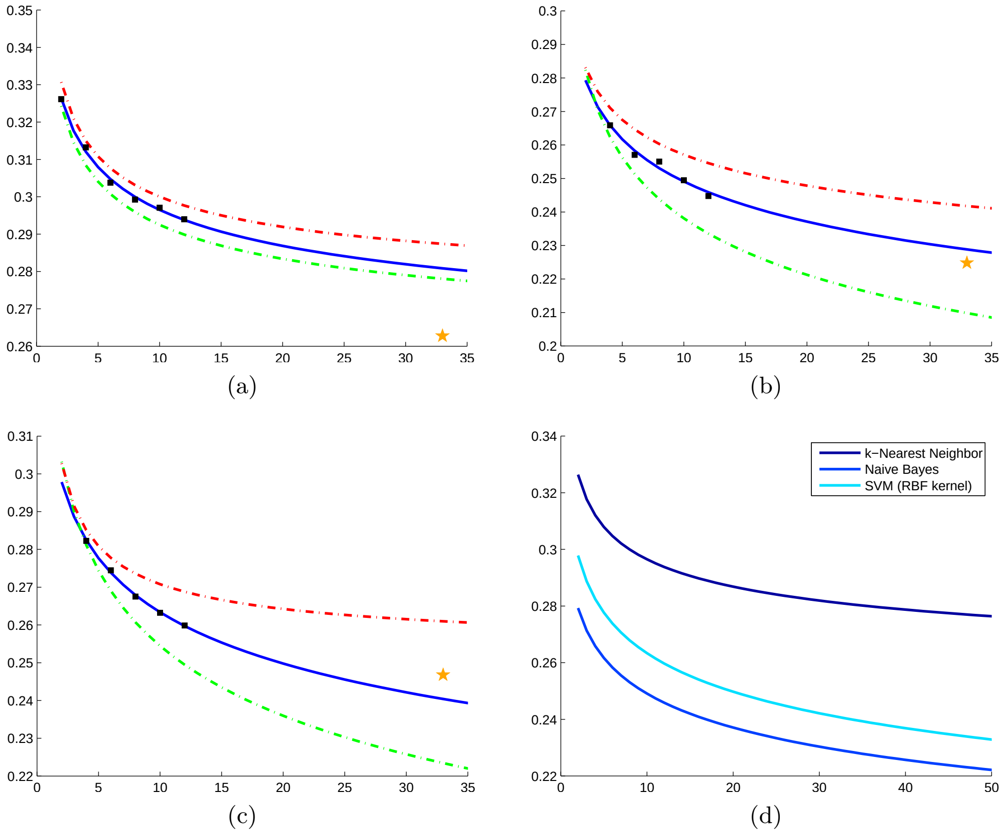
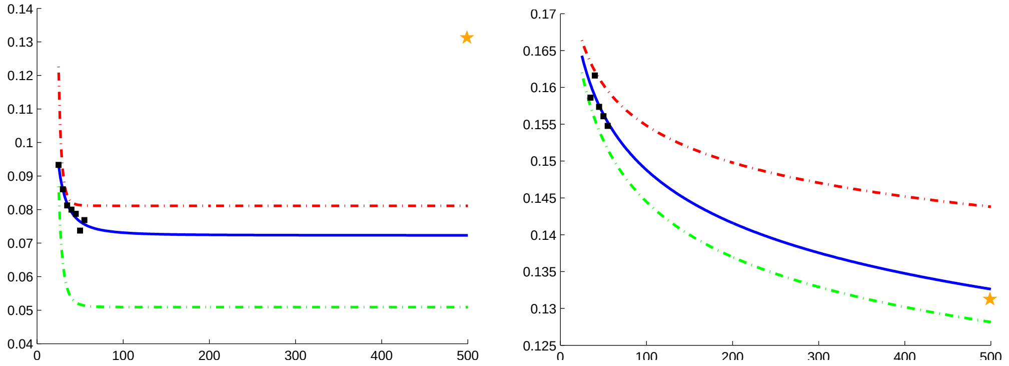


Figure 7.6: Learning curves (blue line) generated for dataset $\mathcal{D}_3$ using mean error rates (black squares) calculated from (a) kNN, (b) NB, and (c) SVM classifiers. Each classifier is accompanied by curves for the 25th (green dashed line) and 75th (red dashed line) percentile of the error as well as LOO error on the validation cohort (yellow star). (d) A direct classifier comparison is made in terms of the mean error rate predicted by each learning curve in (a)-(c).

		n=25	n=30	n=35	n=40	n=45	n=50	n=55	$\overline{\text{IQR}}$
No CV	P25	0.0833	0.0833	0.0417	0.0417	0.0417	0.0417	0.0833	0.0297
	P75	0.1250	0.0833	0.0833	0.0833	0.0833	0.0833	0.0833	
With CV	P25	–	–	0.1563	0.1579	0.1538	0.1514	0.1522	0.0070
	P75	–	–	0.1609	0.1657	0.1618	0.1596	0.1588	

Table 7.4: A comparison between 25th (P25) and 75th (P75) percentile error rates for dataset $\mathcal{D}_1$ using traditional RRS (No CV) and our approach (With CV), with mean interquartile range ($\overline{\text{IQR}}$) shown across all n. Missing values correspond to error rates that did not achieve significance in the permutation test.



(a) (b)

Figure 7.7: Learning curves generated for dataset $\mathcal{D}_1$ using (a) traditional RRS and (b) cross-validated RRS in conjunction with a Naive Bayes classifier. For both figures, mean error rates from the subsampling procedure (black squares) are used to extrapolate learning curves (solid blue line). Corresponding learning curves for 25th (green dashed line) and 75th (red dashed line) percentile of the error are also shown. The error rate from leave-one-out cross-validation is illustrated by a yellow star.

Chapter 8

Experimental design

8.1 Commonly used notation in this chapter

Datasets used in this chapter are defined as $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ (Table 8.1). Note that, although they share a common notation, these datasets are unrelated to the datasets used in Chapter 7.

8.2 Experiment 1: Multi-FOV classification using nuclear architecture

This experiment employs the cohort of 126 ER+ BCa DP slides ($\mathcal{D}_1$) previously presented in Section 6.5.1 (Table 8.1). The goals of this experiment are twofold: (1) multi-FOV classification of whole-slide ER+ BCa DP images and (2) identification of the salient QH features able to distinguish low, intermediate, and high mBR grade at different FOV sizes. First, image features describing nuclear architecture $\mathbf{f}_{\mathrm{NA}}$ are extracted as described in Section 5.2. A multi-FOV classifier is constructed for $\mathbf{f}_{\mathrm{NA}}$ and evaluated in terms of its ability to distinguish patients with low, intermediate, and high mBR grade. Since the multi-FOV classifier (Chapter 6) utilizes a trained classifier, it is susceptible to the arbitrary selection of training and testing data. A 3-fold cross-validation scheme is used to mitigate this bias by splitting the data cohort into 3 subsets in a randomized fashion, from which 2 subsets are used for training and the remaining subset is used for evaluation. The subsets are subsequently rotated until a multi-FOV prediction $\mathbf{H}(\mathbf{D}; \bar{\mathbf{f}}_{\mathrm{NA}})$ is made for each slide. The multi-FOV predictions for all slides are thresholded to create receiver operating characteristic (ROC) curves using the respective mBR grades as ground truth. The entire cross-validation procedure is

	Dataset $\mathcal{D}_1$	Dataset $\mathcal{D}_2$	Dataset $\mathcal{D}_3$
Ground truth	mBR grade	Oncotype DX	tubule subscore from mBR grade
Classes (# samples)	Low (46) Intermed. (60) High (20)	Low (9) Intermed. (11) High (9)	Low (20) Intermed. & High (85)
Staining	H & E	H & E CD34 IHC	H & E
Resolution	1 µm/pixel	1 µm	0.5 µm
Size	whole-slide	whole-slide	500×500 pixel FOVs

Table 8.1: A summary of the datasets used in this chapter.

repeated 20 times, with the mean and standard deviation of the area under the ROC curve (AUC) reported. Note that, since the most relevant FOV sizes are not known *a priori*, we consider a wide range of FOV sizes $T = \{4000, 2000, 1000, 500, 250\}$µm to capture as many variations in tumor morphology as possible.

The inclusion of feature selection in the multi-FOV framework allows us to gain a deeper understanding of the QH features that are most relevant to quantifying tumor morphology and stratifying disease aggressiveness. In this experiment, we rank the mRMR-selected QH features $\bar{\mathbf{f}}_{\text{NA}}$ at each FOV size and explore trends in the selected features across different FOV sizes. In addition, we examine the effect of each additional salient feature on on the cumulative accuracy of the multi-FOV classifier.

8.3 Experiment 2: Multi-FOV classification using nuclear texture

Similar to the procedure outlined in Experiment 1, we employ Dataset $\mathcal{D}_1$ to evaluate the ability of image features characterizing nuclear texture $\mathbf{f}_{\text{NT}}$ to distinguish whole-slide ER+ BCa histopathology based on mBR grade using the multi-FOV framework. Parameter settings, feature selection, and experimental results are reported as described in Section 8.2.

8.4 Experiment 3: Comparison of multi-FOV framework to classification across multiple image resolutions

Although this thesis focuses on the combination of FOVs of different sizes, the ability to integrate image information at various spatial resolutions is also important for the characterization of digitized histopathology slides [45]. For comparison to the multi-FOV approach, a multi-resolution classifier is constructed using Dataset $\mathcal{D}_1$ by re-extracting each FOV of size $\tau = 1000\,\mu\text{m}$ at spatial resolutions of $\kappa \in \{0.25, 0.5, 1, 2, 4\}\mu\text{m/pixel}$. A consensus multi-resolution prediction is achieved for each histopathology slide in a manner analogous to the multi-FOV approach (Chapter 6), except that data is aggregated over all spatial resolutions rather than FOV sizes.

8.5 Experiment 4: Multi-parametric multi-FOV classification using nuclear architecture and microvessel density

This experiment employs Dataset $\mathcal{D}_2$, where each study includes both H & E-stained and CD34 IHC-stained slides. Image features describing nuclear architecture $\mathbf{f}_{\text{NA}}$ and the density of vascular formation f_{VD} are extracted as described in Sections 5.2 and 5.5, respectively. Here, we make the distinction between *global* vascular density (i.e. fraction of DAB-stained pixels in entire slide) from *local* vascular density (i.e. fraction of brown pixels from a smaller FOV (of size $\tau \in T$), the latter of which is used within the multi-FOV framework. Parallel multi-FOV classifiers are constructed for $\mathbf{f}_{\text{NA}}$ and f_{VD} similar to Experiment 1, except for slightly different FOV sizes of $T_{\text{NA}} = \{250, 500, 1000, 2000\}$ and $T_{\text{VD}} = \{250, 500, 1000\}$, respectively. The multi-FOV classifiers $\mathbf{H}(\mathbf{D}; \mathbf{f}_{\text{NA}})$ and $\mathbf{H}(\mathbf{D}; f_{\text{VD}})$ are evaluated both individually and as a fused classifier $\hat{\mathbf{H}}$ (Section 6.4) in terms of their ability to distinguish patients with low, intermediate, and high Oncotype DX Recurrence Scores.

8.6 Experiment 5: Quantifying degree of tubule formation in breast cancer DP

In this experiment, the degree of tubule formation f_{TD} (as defined in Section 5.4) is extracted from each image in Dataset $\mathcal{D}_3$, which comprises 105 FOVs taken from 14 patients. All images were taken from H & E stained BCa DP images digitized at 20x optical magnification (0.5 µm/pixel). For each image, an expert pathologist provided ground truth via the subscore characterizing the extent tubule formation (ranging from 1 to 3) from the mBR grading system [18], where lower scores correspond to well-differentiated tumors with better outcomes and vice versa. In this work, we evaluate the ability of f_{TD} to operate as an effective diagnostic indicator by classifying each image as having either a low (subscore 1) or high (subscore 2 or 3) degree of tubule formation. The ability of f_{TD} distinguish between low (subscore 1) and high (subscores 2 and 3) degrees of tubule formation was evaluated by thresholding f_{TD}. The singular feature f_{TD} is subsequently thresholded to form an ROC curve and an AUC value is reported along with classification accuracy and positive predictive value (PPV) at the ROC operating point.

Chapter 9

Results and discussion

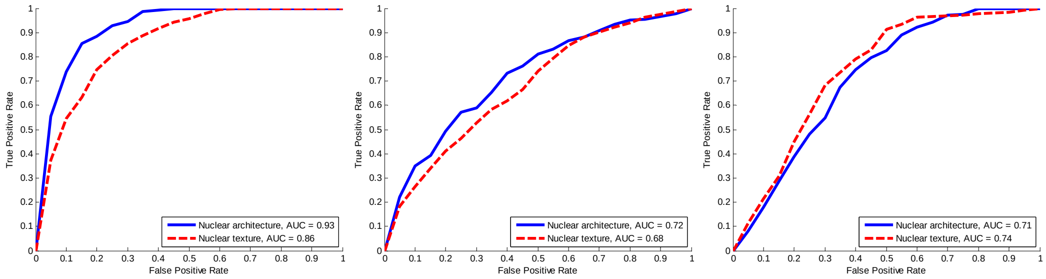
9.1 Commonly used notation in this chapter

The Datasets $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ used in this chapter are the same as the ones defined in Table 8.1 (Chapter 8)). Note that they are different from the datasets considered in Chapter 7.

9.2 Quantitative evaluation and comparison of multi-FOV classification in Experiments 1 and 2

Quantitative results from Experiments 1 (Section 8.2 and 2 (Section 8.3) suggest that predictions made by nuclear architecture $\mathbf{H}(\mathbf{D};\bar{\mathbf{f}}_{\text{NA}})$ and nuclear texture $\mathbf{H}(\mathbf{D};\bar{\mathbf{f}}_{\text{NT}})$ both perform well in characterizing mBR grade in entire ER+ BCa histopathology slides (Figure 9.1). Specifically, nuclear architecture appears to yield higher area under the curve (AUC) values than nuclear texture (AUC of 0.93 and 0.86) in terms of discriminating low vs. high mBR grade. By contrast, both nuclear architecture and nuclear texture yield similar results for distinguishing low vs. intermediate (AUC of 0.72 and 0.68) and intermediate vs. high mBR grade (AUC of 0.71 and 0.74) slides, respectively.

To mitigate the challenges associated with large feature sets (as discussed in Section 1.6.3), the ROC curves in Figure 9.1 were constructed using feature subsets selected by the mRMR algorithm described in Section 5.6. For each experiment, Tables 9.1-9.5 show the features selected at each FOV size along with the cumulative classification accuracy of the multi-FOV approach with the inclusion of each additional feature. Note that some experiments, e.g. nuclear architecture for low vs. high grading (Table 9.1)



(a) (b) (c)

Figure 9.1: Mean receiver operating characteristic (ROC) curves over 20 trials of 3-fold cross-validation for multi-FOV classifiers distinguishing (a) low vs. high mBR grade, (b) low vs. intermediate mBR grade, and (c) intermediate vs. high grade mBR grade. For each task, ROC curves are shown for both nuclear architecture and nuclear texture feature sets along with associated AUC values.

and nuclear texture for intermediate vs. high grading (Table 9.6), demonstrate considerable improvement in classification accuracy with the addition of relevant features while other experiments, e.g. nuclear texture for low vs. intermediate grading (Table 9.5), reach a plateau with the selection of only one or two features.

9.2.1 Trends in salient nuclear architecture features selected across different FOV sizes

In addition to improved classification accuracy, the feature selection process also reveals the specific features that best distinguish low and high grade cancers. For example, Table 9.1 suggests that the average number of neighboring nuclei in a 10 µm radius around each nucleus is the most discriminating feature in smaller FOVs (1000 µm, 500 µm, and 250 µm), but has lesser importance in larger FOV sizes of 2000 µm and 4000 µm, where it is ranked third and fourth, respectively. Conversely, graph-based features derived from the VG and DT appear to play a greater role in larger FOVs, where variations in VG chord length, DT side length, and DT area are more important than nearest neighbor statistics. This pattern is further reinforced in the features selected for distinguishing low vs. intermediate grades (Table 9.2) and intermediate vs. high grades (Table 9.3).

9.2.2 Trends in salient nuclear texture features selected across different FOV sizes

By examining the Haralick co-occurrence features selected for nuclear texture (Tables 9.4-9.6), the dominant role of contrast statistics (especially variance and entropy) is immediately apparent. In addition, the information measure of correlation is shown to have importance for discriminating smaller FOVs ($\tau \in \{250, 500\}$) and data across all three channels (hue, saturation, and intensity) appear to be equally relevant in terms of meaningful feature extraction.

9.3 Experiment 3: comparison between multi-resolution and multi-FOV classifiers

Using all selected features from each classification task (Tables 9.1-9.6), the multi-FOV approach is further evaluated via comparison to a multi-resolution scheme. A comparison of AUC values between the two methods (Table 9.7) suggests that the aggregation of image features at multiple fields-of-view (i.e. multi-FOV classifier) is able to outperform the aggregation of image features at multiple spatial resolutions (ie. multi-resolution classifier) for the grading of BCa histopathology slides. For nuclear architecture $\mathbf{f}_{\mathrm{NA}}$, the superiority of the multi-FOV approach in terms of differentiating low vs. high grades (AUC $= 0.93 \pm 0.012$), low vs. intermediate grades (AUC $= 0.72 \pm 0.037$), and intermediate vs. high grades (AUC $= 0.71 \pm 0.051$) is expected since the spatial arrangement of nuclei is invariant to changes in image resolution. In addition, the ability of a nuclear textural features $\mathbf{f}_{\mathrm{NT}}$ to perform comparably to nuclear architecture in distinguishing low vs. high grades (AUC $= 0.84 \pm 0.036$) and low vs. intermediate grades (AUC $= 0.67 \pm 0.074$) is also unsurprising since textural representations of nuclei will reveal different types of class discriminatory information at various image resolutions. These results suggest that an intelligent combination of the multi-FOV and multi-resolution approaches may yield improved classification of tumor grade in whole-slide BCa histology.

9.4 Experiment 4: multi-parametric multi-FOV classification of H & E and CD34 IHC-stained DP

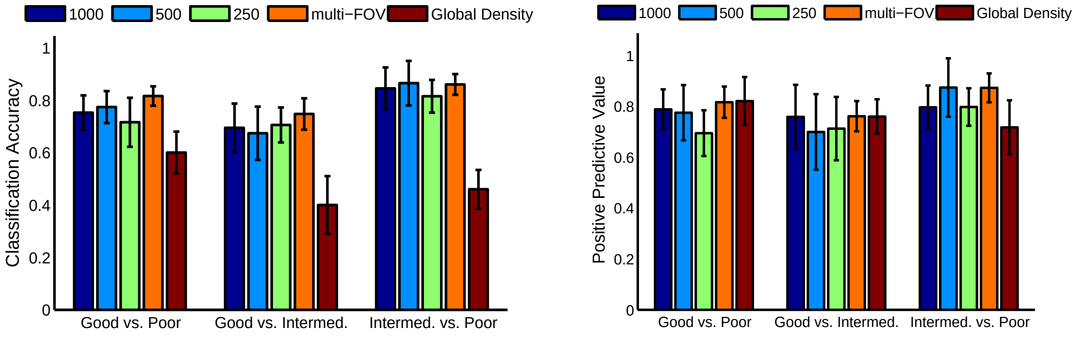
9.4.1 Validation of multi-FOV approach for distinguishing Oncotype DX RS via microvessel density

Similar to the experimental validation used in Section 6.5, the ability of the multi-FOV classifier to outperform classification at individual FOV sizes is borne out by the local vascular density (Figure 9.2), which is able to distinguish entire CD34 IHC stained slides with good vs. poor, good vs. intermediate, and intermediate vs. poor Oncotype DX RS values with classification accuracies of 0.82 ± 0.04, 0.75 ± 0.06, 0.86 ± 0.04, respectively, and positive predictive values (PPV) of 0.82 ± 0.06, 0.76 ± 0.06, 0.87 ± 0.06, respectively. Figure 9.2 demonstrates that multi-FOV classifiers perform as well as (and usually better than) individual FOV sizes in terms of both classification accuracy and PPV. Two-sample t-tests are performed to confirm the significance of this comparison using alternative hypotheses asserting that the multi-FOV classifier outperforms individual FOV sizes in terms of classification accuracy. For good vs. poor outcome, we were able to reject the null hypothesis for all FOV sizes with $p < 0.05$ (Table 9.8). Note that all p-values have been corrected for multiple comparisons using the Bonferroni approach [89].

By contrast, global vascular density produces corresponding classification accuracies of 0.60 ± 0.08, 0.40 ± 0.11, 0.46 ± 0.07 and PPV of 0.82 ± 0.09, 0.76 ± 0.07, and 0.72 ± 0.11, respectively (Figure 9.2), which is consistently worse than the multi-FOV classifier used in conjunction with local vascular density. The superior performance of the multi-FOV classifier is likely due to its ability to capture local variations in vascular density and robustness to intra-slide heterogeneity.

9.4.2 Validation of multi-FOV approach for distinguishing Oncotype DX RS via nuclear architecture

Figure 9.3 shows that the architectural features (in conjunction with the multi-FOV classifier) are able to discriminate H & E stained slides with good vs. poor, good vs.



(a)

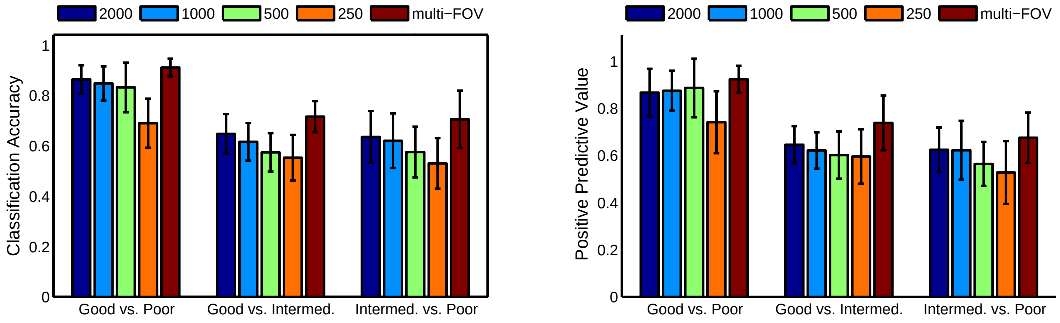
(b)

Figure 9.2: (a) Classification accuracy and (b) positive predictive values for the multi-FOV framework using local vascular density from 29 CD34 IHC stained histopathology slides over 10 trials of 3-fold cross-validation. Note that the bar colors represent different FOV sizes as indicated. For comparison, global vascular density was also calculated directly from each slide and evaluated.

intermediate, and intermediate vs. poor Oncotype DX RS at classification accuracies of 0.91 ± 0.04, 0.72 ± 0.06, 0.71 ± 0.11, respectively, and positive predictive values of 0.92±0.06, 0.74±0.12, 0.68±0.11, respectively. The argument in favor of the multi-FOV classifier is even stronger here than with the IHC-stained images (Section 9.4.1), where it shows significantly increased performance over individual FOV sizes (Figure 9.3). Again, p-values from two-sample t-tests are used to show that the multi-FOV classifier significantly outperforms 3 out of 4 individual FOV sizes with p< 0.05 when comparing good vs. intermediate outcomes and with $p < 0.10$ for 2 of 4 FOV sizes when comparing intermediate vs. poor outcomes.

9.4.3 Evaluation of fused classifier resulting from multi-parametric combination of H & E and IHC-stained DP

Performing a decision-level combination of vascular density and nuclear architecture (Section 6.4) produces classification accuracies of 0.91 ± 0.02, 0.76 ± 0.05, 0.83 ± 0.08 and PPV of 0.94 ± 0.10, 0.85 ± 0.11, 0.92 ± 0.13, for distinguishing good vs. poor, good vs. intermediate, and intermediate vs. poor RS values, respectively (Table 9.9).



(a) (b)

Figure 9.3: (a) Classification accuracy and (b) positive predictive values for the multi-FOV framework using architectural features from 29 H & E stained histopathology slides over 10 trials of 3-fold cross-validation. Note that the bar colors represent different FOV sizes as indicated.

The fact that vascular density and nuclear architecture exploit such disparate aspects of cancer biology (i.e. angiogenesis and tissue morphology, respectively) suggests that the two feature classes are complimentary and integration will yield improved classification. Experiment 3 shows that a decision-level combination of the two feature sets maintains high levels of classification accuracy while improving positive predictive values (Table 9.9) over the corresponding multi-FOV classifiers from Experiments 1 and 2 (Figures 9.2 and 9.3).

9.5 Experiment 5: distinguishing mBR tubule subscore via quantification of tubule density

Experiment 5 (Section 8.6 evaluates the efficacy of our tubule density (defined in Section 5.4) as a diagnostic indicator for BCa histopathology by using it to distinguish between ER+ breast cancers with low (1) and high (2 and 3) mBR tubule subscores. In Figure 9.4, we demonstrate visually that images determined by an expert pathologist to have low and high degrees of tubule formation are clearly separable by f_{TD}. This is further reflected by thresholding f_{TD}, which yields an ROC curve with AUC of 0.94. A classificaton accuracy of 0.89 and positive predictive value of 0.91 were calculated at

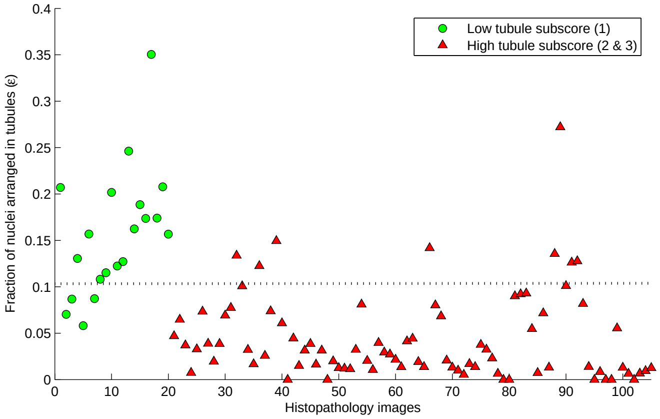


Figure 9.4: A graph showing the fraction of nuclei arranged in tubular formation f_{TD} (y-axis) for each histopathology image (x-axis). The images are arranged by pathologist-assigned scores (i.e. ground truth) denoting degree of tubular formation (ranging from 1-3), which is a component of the mBR grading system. The dotted line represents the operating point ($f_{\mathrm{TD}} = 0.105$) optimally distinguishing low (subscore 1) and high (subscores 2 and 3) degrees of tubule formation with classification accuracy of 0.89 and positive predictive value of 0.91.

the ROC operating point where $f_{\mathrm{TD}} = 0.105$ (Figure 9.5). Our results are also confirmed by an unpaired, two-sample t-test which suggests that images with low and high degrees of tubular formation are indeed drawn from different underlying distributions (i.e. rejects the null hypothesis) with a p-value less than 0.0001.

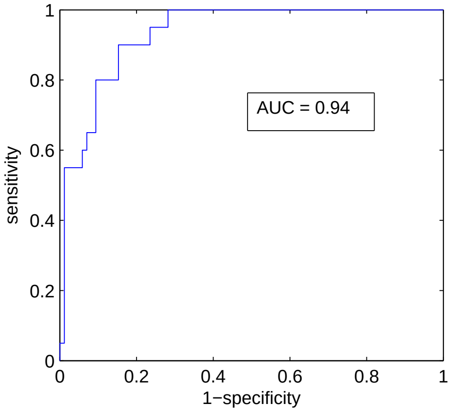


Figure 9.5: A ROC curve generated by thresholding f_{TD} produces an AUC value of 0.94 for distinguishing ER+ breast cancer DP images based on low and high mBR tubule subscores.

Rank	τ	Feature Description	Cum. Acc.
1	4000	VG chord length: min/max ratio	0.85 ± 0.033
	2000	DT side length: min/max ratio	
	1000	# nuclei in 10 µm radius: mean	
	500	# nuclei in 10 µm radius: mean	
	250	# nuclei in 10 µm radius: mean	
2	4000	DT area: min/max ratio	0.86 ± 0.056
	2000	DT area: disorder	
	1000	# nuclei in 10 µm radius: disorder	
	500	DT area: min/max ratio	
	250	Dist. to 5 nearest nuclei: disorder	
3	4000	VG area: min/max ratio	0.88 ± 0.038
	2000	# nuclei in 10 µm radius: mean	
	1000	DT area: min/max ratio	
	500	# nuclei in 10 µm radius: disorder	
	250	DT area: min/max ratio	
4	4000	# nuclei in 10 µm radius: mean	0.91 ± 0.023
	2000	MST edge length: min/max ratio	
	1000	MST edge length: min/max ratio	
	500	DT side length: min/max ratio	
	250	Dist. to 7 nearest nuclei: disorder	
5	4000	VG perimeter: min/max ratio	0.91 ± 0.015
	2000	# nuclei in 10 µm radius: disorder	
	1000	DT side length: min/max ratio	
	500	Dist. to 7 nearest nuclei: disorder	
	250	DT side length: min/max ratio	

Table 9.1: Selected nuclear architecture features at various FOV sizes for low vs. high mBR grade classification.

Rank	τ	Feature Description	Cum. Acc.
1	4000	VG perimeter: min/max ratio	0.71 ± 0.0042
	2000	DT area: disorder	
	1000	DT area: disorder	
	500	# nuclei in 10 µm radius: disorder	
	250	DT side length: min/max ratio	
2	4000	VG chord length: min/max ratio	0.71 ± 0.011
	2000	DT side length: min/max ratio	
	1000	VG chord length: min/max ratio	
	500	Dist. to 7 nearest nuclei: disorder	
	250	# nuclei in 10 µm radius: mean	
3	4000	DT area: disorder	0.73 ± 0.028
	2000	# nuclei in 10 µm radius: mean	
	1000	# nuclei in 10 µm radius: mean	
	500	Dist. to 5 nearest nuclei: disorder	
	250	Dist. to 3 nearest nuclei: disorder	
4	4000	MST edge length: min/max ratio	0.74 ± 0.037
	2000	VG perimeter: min/max ratio	
	1000	VG perimeter: min/max ratio	
	500	DT area: min/max ratio	
	250	# nuclei in 10 µm radius: disorder	

Table 9.2: Selected nuclear architecture features at various FOV sizes for low vs. intermediate mBR grade classification.

Rank	τ	Feature Description	Cum. Acc.
1	4000	VG area: min/max ratio	0.70 ± 0.035
	2000	DT area: disorder	
	1000	DT area: disorder	
	500	VG area: std. dev.	
	250	DT area: min/max ratio	
2	4000	DT area: disorder	0.71 ± 0.054
	2000	VG perimeter: min/max ratio	
	1000	VG chord length: min/max ratio	
	500	# nuclei in 10 µm radius: mean	
	250	Dist. to 7 nearest nuclei: disorder	
3	4000	DT side length: min/max ratio	0.72 ± 0.048
	2000	# nuclei in 10 µm radius: mean	
	1000	MST edge length: min/max ratio	
	500	DT area: disorder	
	250	# nuclei in 40 µm radius: mean	
4	4000	VG chord: min/max ratio	0.73 ± 0.056
	2000	DT side length: min/max ratio	
	1000	DT side length: min/max ratio	
	500	DT area: min/max ratio	
	250	# nuclei in 10 µm radius: mean	

Table 9.3: Selected nuclear architecture features at various FOV sizes for intermediate vs. high mBR grade classification.

Rank	τ	Feature Description	Cum. Acc.
1	4000	Val: Contrast variance - std. dev.	0.80 ± 0.047
	2000	Hue: Contrast variance - mean	
	1000	Sat: Contrast variance - std. dev.	
	500	Val: Contrast variance - std. dev.	
	250	Val: Contrast entropy - disorder	
2	4000	Sat: Contrast variance - std. dev.	0.81 ± 0.044
	2000	Sat: Contrast variance - mean	
	1000	Hue: Contrast variance - mean	
	500	Hue: Info. measure 1 - std. dev.	
	250	Hue: Info. measure 1 - std. dev.	
3	4000	Hue: Contrast variance - std. dev.	0.84 ± 0.040
	2000	Hue: Contrast variance - std. dev.	
	1000	Val: Contrast variance - std. dev.	
	500	Val: Contrast entropy - disorder	
	250	Val: Contrast average - std. dev.	

Table 9.4: Selected nuclear texture features at various FOV sizes for low vs. high mBR grade classification.

Rank	τ	Feature Description	Cum. Acc.
1	4000	Hue: Contrast variance - disorder	0.69 ± 0.024
	2000	Sat: Contrast variance - mean	
	1000	Val: Contrast average - std. dev.	
	500	Sat: Contrast variance - std. dev.	
	250	Sat: Info. measure 1 - std. dev.	
2	4000	Val: Contrast variance - std. dev.	0.69 ± 0.027
	2000	Val: Contrast variance - std. dev.	
	1000	Sat: Contrast inv. moment - std. dev.	
	500	Sat: Info. measure 1 - std. dev.	
	250	Sat: Contrast variance - std. dev.	
3	4000	Val: Contrast entropy - disorder	0.70 ± 0.024
	2000	Sat: Contrast average - std. dev.	
	1000	Hue: Intensity average - disorder	
	500	Val: Info. measure 1 - std. dev.	
	250	Sat: Contrast inv. moment - std. dev.	

Table 9.5: Selected nuclear texture features at various FOV sizes for low vs. intermediate mBR grade classification.

Rank	τ	Feature Description	Cum. Acc.
1	4000	Hue: Contrast variance - std. dev.	0.68 ± 0.082
	2000	Hue: Contrast variance - mean	
	1000	Sat: Contrast variance - std. dev.	
	500	Val: Contrast variance - std. dev.	
	250	Val: Contrast variance - std. dev.	
2	4000	Hue: Contrast variance - mean	0.75 ± 0.044
	2000	Val: Contrast entropy - disorder	
	1000	Hue: Contrast variance - mean	
	500	Sat: Contrast variance - std. dev.	
	250	Sat: Contrast variance - std. dev.	
3	4000	Sat: Contrast variance - mean	0.74 ± 0.040
	2000	Hue: Contrast variance - std. dev.	
	1000	Hue: Contrast variance - std. dev.	
	500	Val: Contrast entropy - disorder	
	250	Hue: Info. measure 1 - std. dev.	
4	4000	Sat: Contrast variance - std. dev.	0.74 ± 0.030
	2000	Sat: Contrast variance - mean	
	1000	Val: Contrast variance - std. dev.	
	500	Sat: Contrast inv. moment - std. dev.	
	250	Sat: Contrast inv. moment - std. dev.	
5	4000	Hue: Contrast variance - disorder	0.75 ± 0.035
	2000	Sat: Contrast variance - std. dev.	
	1000	Sat: Contrast variance - mean	
	500	Val: Entropy - std. dev.	
	250	Val: Contrast entropy - disorder	

Table 9.6: Selected nuclear texture features at various FOV sizes for intermediate vs. high mBR grade classification.

Experiment	Feature Set	Multi-FOV	Multi-Res.
Low vs. high	$\bar{\mathbf{f}}_{\mathrm{NA}}$	0.93 ± 0.012	0.86 ± 0.035
	$\bar{\mathbf{f}}_{\mathrm{NT}}$	0.86 ± 0.036	0.84 ± 0.036
Low vs. intermed.	$\bar{\mathbf{f}}_{\mathrm{NA}}$	0.72 ± 0.037	0.67 ± 0.049
	$\bar{\mathbf{f}}_{\mathrm{NT}}$	0.68 ± 0.028	0.67 ± 0.074
Intermed. vs. high	$\bar{\mathbf{f}}_{\mathrm{NA}}$	0.71 ± 0.051	0.65 ± 0.054
	$\bar{\mathbf{f}}_{\mathrm{NT}}$	0.74 ± 0.036	0.66 ± 0.075

Table 9.7: Area under the ROC curve (AUC) values for the comparison of low, intermediate, and high grade cancers using both multi-FOV and multi-resolution classifiers.

FOV size	Good vs. Poor	Good vs. Intermed.	Intermed. vs. Poor
	Vascular density in IHC stained histopathology		
1000	0.0288	0.2250	0.9042
500	0.0123	0.1011	1.0000
250	0.0129	0.2313	0.1101
	Nuclear architecture in H & E stained histopathology		
2000	0.0570	0.0666	1.0000
1000	0.0267	0.0066	0.1575
500	0.0429	0.0003	0.0657
250	<0.0001	<0.0001	0.0027

Table 9.8: Bonferroni-corrected *p*-values produced by two-sided t-tests with a null hypothesis that classification results from the multi-FOV approach are equivalent to results from individual FOV sizes from both IHC stained and H & E stained histopathology slides. The alternative hypothesis asserts that the multi-FOV classifier performs better than individual FOV sizes.

	Good vs. Poor	Good vs. Intermed.	Intermed. vs. Poor
Accuracy	0.91 ± 0.022	0.76 ± 0.051	0.83 ± 0.076
PPV	0.94 ± 0.10	0.85 ± 0.11	0.92 ± 0.13

Table 9.9: Classification accuracies and positive predictive values (PPV) for comparing good, intermediate, and poor Oncotype DX scores via the multi-FOV framework using a combination of vascular density and architectural features over 10 trials of 3-fold cross-validation.

Chapter 10

Concluding Remarks

In this thesis we have presented a QH-based companion diagnostic framework containing tools for the quantitative prediction of disease outcome in early stage, ER+ BCa patients using only QH features extracted from whole-slide H & E stained DP images. Specific goals accomplished in this work include:

- Color standardization of H & E stained DP images by accounting for differing proportions of tissue structures,
- Detection of tubule formation in ER+ BCa histopathology by using O'Callaghan neighborhoods to enforce domain constraints between nuclei and lumen followed by the definition of tubule density as a QH descriptor,
- A robust multi-FOV framework for sampling and combining class predictions across FOVs at different sizes, and
- Robust method for predicting large-scale classifier performance and performing classifier comparison studies using limited training data.

The following paragraphs detail the major findings for each tasks considered in this work.

First, we present a EM-based segmentation-driven standardization (EMS) algorithms employs an independent and localized approach to correcting nonstandardness that accounts for the varying proportions of different tissue classes (e.g. epithelium, stroma, nuclei) in H & E stained DP. EMS will enable the creation of more robust object detection and segmentation methods, which are becoming increasingly elaborate and time-consuming in an effort to account for the highly variable appearance of tumor morphology in H & E stained DP. To this end, we performed a segmentation of nuclei in H & E stained DP images simply by thresholding the intensity channel and showed that images corrected by EMS yield a more consistent segmentation result than both unstandardized data and a more naive global standardization technique. An additional benefit of EMS over other approaches is its ability to operate on retrospective data where information about staining process or scanning systems is unavailable.

Improved detection and segmentation of low-level tissue objects (e.g. nuclei, lumen) will allow for the identification of more complex histological structures such as glands and lumen. In this work, we presented a method for the systemic incorporation of domain knowledge via the O'Callaghan neighborhood to identify tubules by constraining the spatial relationship between two sets of low-level objects (nuclei and lumen). In addition, a novel feature set comprising 22 O'Callaghan neighborhood-based descriptors was created to distinguish tubules from confounding structures in BCa DP images. The accurate delineation of tubules led to the subsequent definition and extraction of *tubule density*, a QH descriptor which was shown to be a good predictor of tubule subscore in mBR grading.

The majority of QH features developed for histopathological imagery are designed for operation on small FOVs. Instead of random or arbitrary sampling of FOVs from whole-slide DP, we presented a multi-FOV framework that employs a multitude of FOVs at various FOV sizes, thus eliminating the need for *a priori* determination of an optimal FOV size. The superiority of this method over classification at individual FOV sizes was demonstrated both theoretically and experimentally in the context of distinguishing low, intermediate, and high risk ER+ BCa patients. Furthermore, the ability to incorporate feature selection into the mult-FOV framework allowed us to gain deeper insight into the specific aspects of tumor morphology that are related to patient outcome. Additionally, we showed the extensibility of the multi-FOV framework to include complementary information from other histopatholgical sources by creating a fused predictor that combined: (1) nuclear architecture descriptors extracted from H & E stained DP and (2) vascular density markers extracted from CD34 IHC-stained DP.

The selection of an appropriate classifier for QH-based companion diagnostics systems is crucial, especially in the context of clinical trials where the classifier must be selected *a priori* with the limited availability of training data. Existing approaches for extrapolating classifier performance from small datasets to larger cohorts, e.g. repeated random sampling (RRS), may suffer from high variability due to the heterogeneous nature of biomedical imaging data. In this work, we presented an extension to RRS that was shown to increase the robustness (i.e. generalizability) of predicted classifier performance by using cross-validation sampling to ensure that all samples are used for both training and testing the classifiers. We were also able to demonstrate an extension of our approach to pixel- and voxel-level studies where data from both classes is found within each patient study, a concept that has previously been unexplored in this regard.

Chapter 11

Future Work

The research presented in this thesis yields numerous paths for future work. In the context of the work presented in this thesis, the logical next steps would involve evaluation against a large cohort of ER+ breast cancer patients accompanied by long-term survival data (i.e. true patient outcome).

More generally, the QH-based companion diagnostic system for ER+ breast cancers presented in this work can be augmented in a number of ways. For instance, the multi-FOV framework currently operates on regions of invasive ductal carcinoma (IDC) as identified by an expert pathologist. Another example is the detection of mitotic nuclei (an important component of mBR grading) in H & E stained histology, a task which has proven to be extremely challenging for computerized algorithms. Automating such tasks would greatly enhance the accuracy, efficiency, and robustness of personalized diagnostics for breast cancer patients.

While the primary application of the multi-FOV framework is for the outcome prediction of ER+ breast cancer patients, the methods developed are generalizable to other fields of interest that require a single class prediction to be achieved through analysis of large heterogeneous images. In terms of DP, the framework can potentially be extended to (1) tissue from any organ of interest and (2) any number of staining protocols or multispectral images containing prognostic information. Other potential extensions to the multi-FOV framework include intelligently learning which FOV sizes are more important for predicting patient outcome and weighting them appropriately.

On a larger scale, the relationship between quantitative imaging signatures extracted from the analysis of histological and radiological imagery is currently being explored in the nascent field of radiohistomorphometrics. In future work, correlating and integrating the salient QH features identified in this work with prognostically-relevant radiological imaging signatures may yield new insights into predicting disease outcome and lead to improved personalized diagnostic solutions.

Chapter 12

Appendices

Appendix A: Glossary of notation, symbols, and abbreviations commonly used in this thesis

DP	digital pathology
QH	quantitative histomorphometry
CAD	computer-aided diagnosis
BCa	breast cancer
ER+	estrogen receptor-positive
LN-	lymph node-negative
IDC	invasive ductal carcinoma
mBR	modified Bloom-Richardson (grading system)
ODX	Oncotype DX (genomic assay)
RS	(Oncotype DX) Recurrence Score
H & E	hematoxylin and eosin (staining)
IHC	immunohistochemical (staining)
FOV	field-of-view
ROC	receiver operating characteristic (curve)
EM	Expectation-Maximization
RGB	red-green-blue (color space)
HSI	hue-saturation-intensity (color space)
AUC	area under the ROC curve
EMS	EM-based standardization
GS	global standardization
NMI	normalized median intensity

GAC	geodesic active contour
CGAC	color gradient-based geodesic active contour
DR	dimensionality reduction
NLDR	non-linear dimensionality reduction
MRI	magnetic resonance imaging
MRS	magnetic resonance spectroscopy
GE	graph embedding
mRMR	minimum redundancy maximum relevance (feature selection)
VG	Voronoi graph
DT	Delaunay triangulation (graph)
MST	minimum spanning tree (graph)
RRS	repeated random sampling
kNN	k-nearest neighbor (classifier)
NB	naive Bayes (classifier)
SVM	Support Vector Machine (classifier)
LOO	leave-one-out (cross-validation)
IQR	interquartile range

$\mathcal{C}$	Image scene
C	2D set of pixels in $\mathcal{C}$
c	Single pixel in C
$g(c)$	Function assigning single color channel value to c
$\mathbf{g}(c) \in \mathbb{R}^3$	Function assigning 3-dimensional color values (e.g. red-green-blue) to c
$\mathcal{Y}(\mathcal{C})$	Class label of sample $\mathcal{C}$
f	Image feature for a sample
$\mathbf{f}$	Set of image features for a sample

Appendix B: Description of three exemplar classifiers

This appendix comprises the three fundamentally different classifiers (k-nearest neighbor (kNN), Naive Bayes (NB), and Support Vector Machine (SVM)) primarily used in Chapter 7. For all methods described below, let us define a training sample $A \in \mathbf{A}$ and testing sample $B \in \mathbf{B}$ with corresponding feature sets $\mathbf{F}(A)$, $\mathbf{F}(B)$ and ground truth labels $\mathcal{Y}(A)$, $\mathcal{Y}(B)$.

k-Nearest Neighbor Classifier

For each testing sample, the kNN classifier identifies the nearest training sample

$$\hat{A}_1 = \operatorname*{argmin}_{A} \mathbb{D}(\mathbf{F}(A), \mathbf{F}(B)), \tag{12.1}$$

where $\mathbb{D}(\cdot, \cdot)$ is a user-specified distance metric. This process is repeated until the k nearest training samples $\{\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_k\}$ have been identified. The class prediction is defined by majority voting across the ground truth labels $\{\mathcal{Y}(\hat{A}_1), \mathcal{Y}(\hat{A}_2), \ldots, \mathcal{Y}(\hat{A}_k)\}$ for all k training samples.

Naive Bayes Classifier

The naive Bayes classifier is a simple approach to statistical inference that relies on the application of Bayes' theorem under the assumptions that (1) a sufficient amount of training data is available and (2) its constituent features are independent [41]. For binary classification, let us define the likelihood of observing class ω_1 given feature set $\mathbf{F}$ as

$$P(\omega_1|\mathbf{F}) = \frac{P(\omega_1)p(\mathbf{F}|\omega_1)}{P(\omega_1)p(\mathbf{F}|\omega_1) + P(\omega_2)p(\mathbf{F}|\omega_2)}, \tag{12.2}$$

where $P(\omega_1)$, $P(\omega_2)$ are the prior probabilities of occurrence of the two classes, and $p(\mathbf{F}|\omega_1)$, $p(\mathbf{F}|\omega_2)$ are the *a priori* class conditional distributions of $\mathbf{F}$. Using all training set samples $\{A : A \in \mathbf{A}, \mathcal{Y}(A) = \omega_1\}$ in class ω_1, *a priori* distributions $p(F_i(\mathbf{A})|\omega_1)$ are generated for each feature $F_i \in \mathbf{F}$. Due to the relatively small amount of training data used in this work, kernel density estimation (KDE) is employed to ensure that each $p(F_i(\mathbf{A})|\omega_1)$ is smooth and continuous [111]. Assuming independence between the features allows the distributions to be collapsed such that $p(\mathbf{F}(\mathbf{A})|\omega_1) = \prod_i p(F_i(\mathbf{A})|\omega_1)$. The *a priori* probability $p(\mathbf{F}(\mathbf{A})|\omega_2)$ is similarly generated using samples $\{A : A \in \mathbf{A}, \mathcal{Y}(A) = \omega_2\}$. Testing sample B is said to be correctly classified if the maximum *a posteriori* decision is equal to the ground truth label, i.e. $\mathcal{Y}(B) = \text{argmax}_{\omega \in \{\omega_1, \omega_2\}} P(\omega) p(\mathbf{F}(B)|\omega)$.

Support Vector Machine Classifier

The SVM classifier operates by projecting training data onto a higher-dimensional space and constructing a hyperplane to maximize the distance between marginal samples in the two object classes [112]. Evaluation is subsequently performed by projecting a testing sample into the same space and ascertaining its location relative to the hyperplane. In this paper, the projection is defined by calculating the radial basis function (RBF) kernel

$$\Pi(A_1, A_2) = e^{\rho \|\mathbf{F}(A_1) - \mathbf{F}(A_2)\|_2^2} \tag{12.3}$$

between all pairs of training samples $A_1, A_2 \in \mathbf{A}$, where ρ is a user-defined scaling parameter. The general form of the SVM prediction function is

$$\Theta(B) = \sum_{\gamma=1}^{\tau} \xi_\gamma \mathcal{Y}(A_\gamma) \Pi(B, A_\gamma) + \mathbf{b}, \tag{12.4}$$

where $A_\gamma \in \mathbf{A}$ represents a marginal training sample (i.e. support vector), $\mathbf{b}$ is the hyperplane bias estimated over all τ support vectors, and ξ_γ is the slack variable that governs the tradeoff between minimizing training error and maximizing margin [112]. The output of the SVM classifier $\Theta(B)$ represents the distance from testing sample B to the hyperplane, which is determined to be classified correctly if $\mathcal{Y}(B) = \text{sign}\,[\Theta(B)]$.

Appendix C: Manifold learning approach to stratification of ER+ breast cancers with good and poor outcomes

During the course of this thesis, a different type of QH-based decision support system was considered for distinguishing patients with low vs. high mBR grade and low vs. high Oncotype DX RS. This system involves (1) manual selection of a single representative FOV from a whole-slide H & E stained DP image, (2) detection of individual epithelial nuclei in the FOV, (3) extraction of graph-based QH features quantifying nuclear architecture, (4) reduction of the feature space via dimensionality reduction, and (5) classification into either low or high RS classes. While this approach is effective for stratifying patients based on disease outcome, it was ultimately deemed unfeasible for outcome prediction for the following reasons.

- Manual FOV selection: This system requires a single representative FOV to be selected from the entire DP slide. This is undesirable for a large-scale analysis because it requires manual intervention and does not account for heterogeneity in the tumor morphology throughout the slide.
- Dimensionality reduction: Unsupervised dimensionality reduction techniques can reveal the underlying manifold of a data set; however, they are highly dependent on the data used and their shape may change dramatically when data is either added or removed. Additionally, the inability to ascertain exactly which features are used to generate the low-dimensional manifold makes it difficult gain an appreciation of the biological process driving the changes in tumor morphology.

Notation used in this appendix

Notation common throughout this thesis can be found in Appendix A. Other symbols used in this appendix are independent of the notation used in the rest of the thesis.

Nuclear detection and graph-based features extraction

In this work, detection of individual nuclei is performed by first dividing an image into four tissue classes via the Expectation-Maximization (EM) algorithm (as described in

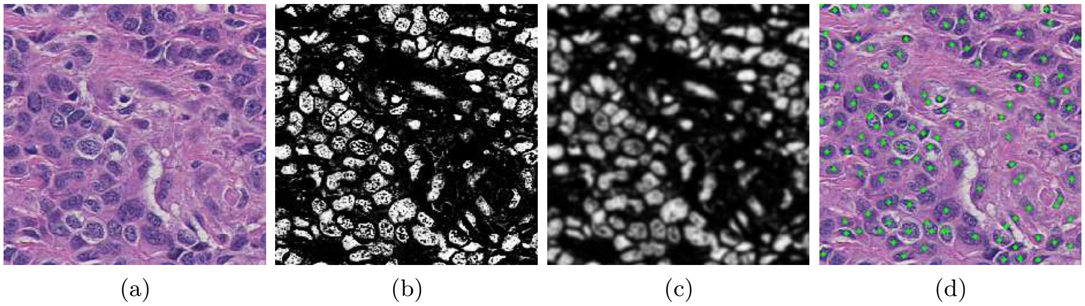


Figure 12.1: (a) An ER+ BCa DP image shown along with corresponding (b) EM-based segmentation of epithelial nuclei. (c) The segmentation is subsequently smoothed and (d) the centroids of individual nuclei are identified by morphological and connected component operations.

Section 3.2). The EM component that best represents epithelial nuclei (Figure 12.1(b)) is selected manually and smoothed (Figure 12.1(c)) to reduce intra-nuclear intensity variations. Morphological and connected component operations are then applied to identify individual objects corresponding to nuclei and the corresponding set of nuclear centroids is found for each image (Figure 12.1(d)).

Detected nuclear centroids are then used for the construction of Delaunay triangulation (Figure 12.2(b), (e)) and minimum spanning tree (Figure 12.2(c), (f)) graphs as described in Sections 5.2.2 and 5.2.3, respectively. A total of 12 features quantifying nuclear architecture $\mathbf{f}(\mathcal{C})$ are extracted from each image $\mathcal{C}$ as shown in Table 5.1.

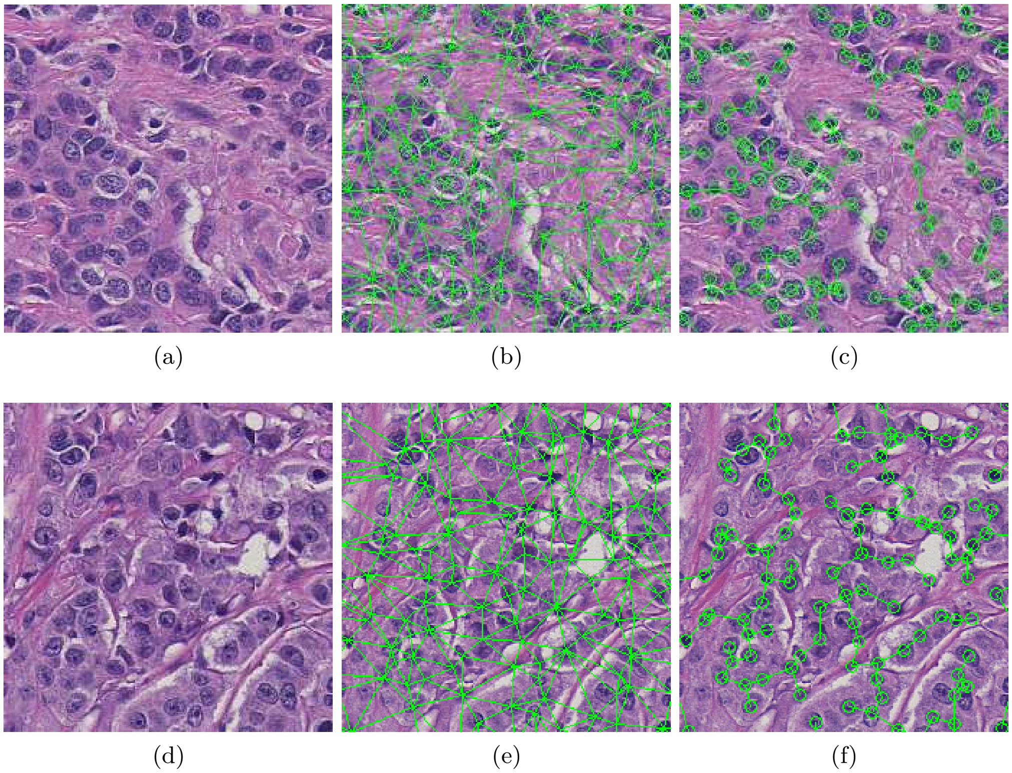
Dimensionality reduction via graph embedding

We use graph embedding (GE) to transform the nuclear architecture feature set into a low-dimensional embedding [30].

Given images $\mathcal{C}_a$ and $\mathcal{C}_b$, a confusion matrix $\mathcal{W}(a,b) = \exp(-\|\mathbf{f}(\mathcal{C}_a) - \mathbf{f}(\mathcal{C}_b)\|_2) \in \mathbb{R}^{N \times N}$, where N is the total number of images, is first constructed $\forall a, b \in \{1, 2, \ldots, N\}$. The optimal embedding vector $\mathbf{f}'$ is obtained from the maximization of the function,

$$\mathcal{E}(\mathbf{f}') = 2(N-1) \cdot \text{trace}\left[\frac{\mathbf{f}'^{\mathsf{T}}(\mathcal{A} - \mathcal{W})\mathbf{f}'}{\mathbf{f}'^{\mathsf{T}}\mathcal{A}\mathbf{f}'}\right],$$

where $\mathcal{A}(a,a) = \sum_b \mathcal{W}(a,b)$. The low-dimensional embedding space is defined by the



(a) (b) (c)

(d) (e) (f)

Figure 12.2: (a), (d) Low and high grade ER+ BCa samples are shown with corresponding (b), (e) Delaunay triangulation and (c), (f) minimum spanning tree graphs overlaid.

eigenvectors corresponding to the β smallest eigenvalues of $(\mathcal{A} - \mathcal{W})\mathbf{f}' = \lambda\mathcal{A}\mathbf{f}'$. Specifically, reducing the high-dimensional feature space to a three-dimensional (3D) sub-space allows us to evaluate (both quantitatively and qualitatively) the discriminability of the image-derived features in distinguishing samples with different cancer grade patterns and hence different prognoses.

Evaluation via a Support Machine Vector classifier in conjunction with randomized cross-validation

A support vector machine (SVM) classifier [112] is constructed as described in Appendix B (with the exception of using a linear kernel [41] rather than the RBF kernel). Training and testing samples are selected via randomized K-fold cross-validation algorithm, whereby the dataset is divided randomly into K subsets. The samples from $K - 1$ subsets are used for training and the remaining subset is used for evaluation.

For each of the K folds, the subsets are rotate to ensure that each sample is classified once. The entire cross-validation scheme is repeated for 100 trials, over which mean and standard deviation of classification accuracy are reported.

Geodesic distance-based projection from 3D to 1D

The 3D GE manifold can be "unwrapped" into a 1D (linear) space simply by selecting the image $\mathcal{C}_1$ at one end of the manifold as an anchor point and using the Euclidean distance metric to find the next nearest image on the 3D manifold. By using $\mathcal{C}_a$ as the new anchor point, this process is repeated until all images have been included. Thus the geodesic distances between all images $\mathcal{C}$ embedded on the manifold are determined and GE is again employed to project the data down to a 1D line. By uncovering the grade (outcome) labels of the samples on this 1D projection and their relative locations, an image-based recurrence score can be determined to distinguish between low, intermediate, and high BCa grades (and hence outcomes). For any new image $\mathcal{C}_b$ projected onto this line, the relative distance of $\mathcal{C}_b$ from poor, intermediate, and good outcome samples on the trained manifold will enable prediction of prognosis for $\mathcal{C}_b$.

Experimental results and discussion

Dataset

A total of 37 H & E stained breast histopathology images were collected from a cohort of 17 patients and scanned into a computer using a high resolution whole slide scanner at 20x optical magnification. Each image is accompanied by a corresponding Oncotype DX Recurrence Score and mBR grade as determined by an expert pathologist.

Distinguishing low vs. high mBR grade

SVM classifiers trained via 100 trials of 3-fold cross-validation on the original $\mathbf{f}$ and reduced (3D) $\mathbf{f}'$ feature sets were able to distinguish high and low grade BCa histopathology images with classification accuracies of 0.75 ± 0.06 and 0.84 ± 0.05, respectively (Table 12.1). These results suggest that GE has embedded the original feature set

Dataset	Ground truth	Automated Detection	Manual Detection
$\mathbf{f}'$	RS	0.84 ± 0.03	0.72 ± 0.05
	Grade	0.84 ± 0.05	0.86 ± 0.05
$\mathbf{f}$	RS	0.85 ± 0.03	0.72 ± 0.06
	Grade	0.75 ± 0.06	0.85 ± 0.05

Table 12.1: Mean and standard deviation of classification accuracy are reported across 100 trials of 3-fold cross-validation evaluating the ability of the original $\mathbf{f}$ and low-dimensional $\mathbf{f}'$ feature sets to distinguish between both Oncotype DX Recurrence Score (RS) and mBR grade labels. Results are presented for experiments using both automatically and manually delineated BCa nuclei.

without any significant loss of information. The success of the architectural features is confirmed qualitatively by the clear separation between high and low BC grade on the 3D manifold (Figure 12.3(a)).

Distinguishing low vs. high Oncotype DX RS

Replacing the grade labels with the RS labels, the SVM trained via 3-fold cross-validation on $\mathbf{f}$ and $\mathbf{f}'$ yielded classification accuracies of 0.85 ± 0.03 and 0.84 ± 0.03, respectively (Table 12.1). This suggests the existence of a relationship between molecular prognostic assays such as Oncotype DX and the spatial arrangement of histological structures in BCa histopathology. The 3D manifolds in Figures 12.3(a), (b) reveal a similar underlying biological stratification that exists in mBR grade and Oncotype DX RS, in turn suggesting that the QH features employed to characterize mBR grade could recapitulate the prognostic capabilities of Oncotype DX. The curvilinear 3D manifold on which the different BC grades (low to high) reside in a smooth continuum that may potentially offer insight into BCa biology as well.

Creating an image-based assay using 1D projection

Figures 12.3(c), (d) represent the 1D projections of the 3D manifolds shown in Figures 12.3(a), (b), respectively. The manifolds reveal a smooth, continuous progression from low to medium to high levels in terms of both RS and histological (grade) for all ER+ BCa samples considered. The similarity between the 1D manifolds (Figures

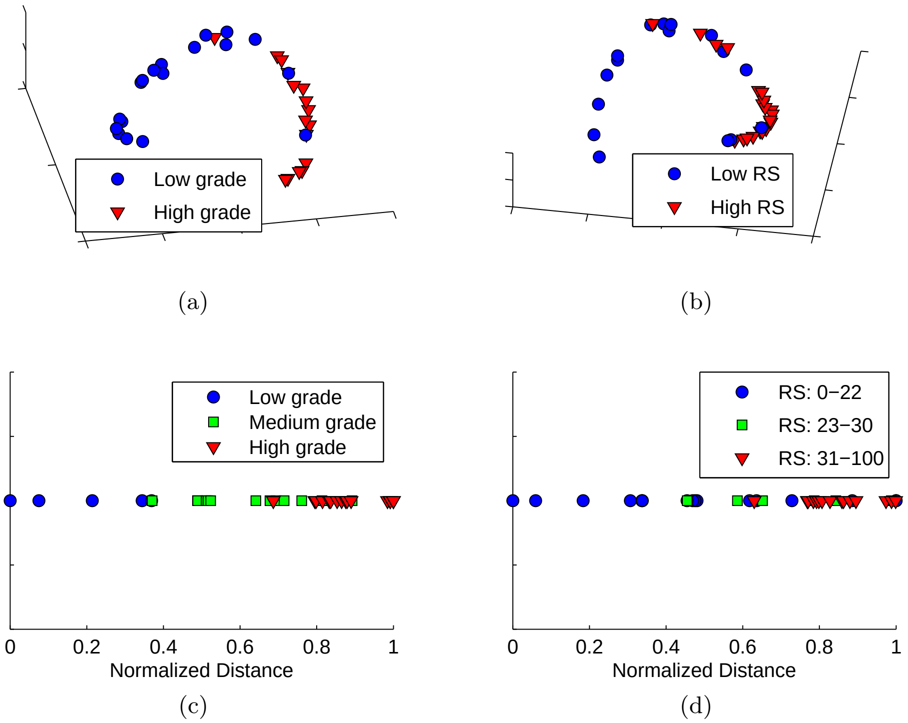


Figure 12.3: Graph Embedding plots of architectural features show clear separation of different (a) BC grades and (b) RS labels. The embeddings are projected into a 1D line, where (c) mBR grade and (d) RS are characterized by a single score.

12.3(c), (d)) suggest that our QH-based approach can be used to generate a prognostic assay to predict survival scores in much the same way as Oncotype DX.

References

[1] Ajay Nagesh Basavanhally, Shridar Ganesan, Shannon Agner, James Peter Monaco, Michael D Feldman, John E Tomaszewski, Gyan Bhanot, and Anant Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Trans. Biomed. Eng.*, 57(3):642–653, Mar 2010.

[2] A Basavanhally, M Feldman, N Shih, C Mies, JE Tomaszewski, S Ganesan, and A Madabhushi. Multi-field-of-view strategy for image-based outcome prediction of multi-parametric estrogen receptor-positive breast cancer histopathology: Comparison to oncotype dx. *J Pathol Inform*, 2, 01/2012 2011.

[3] Ajay Basavanhally, Shridar Ganesan, Michael Feldman, Natalie Shih, Carolyn Mies, John Tomaszewski, and Anant Madabhushi. Multi-field-of-view framework for distinguishing tumor grade in er+ breast cancer from entire histopathology slides. *IEEE Trans. Biomed. Eng.*, 2013.

[4] A. Basavanhally, Jun Xu, A. Madabhushi, and S. Ganesan. Computer-aided prognosis of er+ breast cancer histopathology and correlating survival outcome with oncotype dx assay. In *Proc. IEEE Int. Symp. Biomedical Imaging: From Nano to Macro ISBI '09*, pages 851–854, 2009.

[5] Ajay Basavanhally, Scott Doyle, and Anant Madabhushi. Predicting classifier performance with a small training set: Applications to computer-aided diagnosis and prognosis. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pages 229–232. IEEE, 2010.

[6] A Basavanhally, E Yu, J Xu, S Ganesan, M Feldman, JE Tomaszewski, and A Madabhushi. Incorporating domain knowledge for tubule detection in breast histopathology using o'callaghan neighborhoods. In *SPIE Medical Imaging*, volume 7963 (1) of *Computer-Aided Diagnosis*, page 796310. SPIE, SPIE, 2011.

[7] Ajay Basavanhally, Shridar Ganesan, Natalie Shih, Carolyn Mies, Michael Feldman, John Tomaszewski, and Anant Madabhushi. A boosted classifier for integrating multiple fields of view: Breast cancer grading in histopathology. In *Biomedical Imaging: From Nano to Macro, IEEE International Symposium on*, pages 125–128, 2011 Mar 30 2011.

[8] Ajay Basavanhally and Anant Madabhushi. Em-based segmentation-driven color standardization of digitized histopathology. In *SPIE Medical Imaging*, pages 86760G–86760G. International Society for Optics and Photonics, 2013.

[9] Gian Kayser and Klaus Kayser. Quantitative pathology in virtual microscopy: history, applications, perspectives. *Acta Histochem*, 115(6):527–532, Jul 2013.

[10] Ronald S. Weinstein, Anna R. Graham, Lynne C. Richter, Gail P. Barker, Elizabeth A. Krupinski, Ana Maria Lopez, Kristine A. Erps, Achyut K. Bhattacharyya, Yukako Yagi, and John R. Gilbertson. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. *Hum Pathol*, 40(8):1057–1069, Aug 2009.

[11] Metin N Gurcan, Laura Boucheron, Ali Can, Anant Madabhushi, Nasir Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Rev Biomed Eng*, 2:147–171, 2009.

[12] Melina B. Flanagan, David J. Dabbs, Adam M. Brufsky, Sushil Beriwal, and Rohit Bhargava. Histopathologic variables predict oncotype dx recurrence score. *Mod Pathol*, 21(10):1255–1261, Oct 2008.

[13] Britta Weigelt and Jorge S. Reis-Filho. Molecular profiling currently offers no more than tumour morphology and basic immunohistochemistry. *Breast Cancer Res*, 12 Suppl 4:S5, 2010.

[14] K. H. Allison, P. L. Kandalaft, C. M. Sitlani, S. M. Dintzis, and A. M. Gown. Routine pathologic parameters can predict oncotype dx recurrence scores in subsets of er positive patients: who does not always need testing? *Breast Cancer Res Treat*, 131(2):413–424, Jan 2012.

[15] Rebecca Siegel, Deepa Naishadham, and Ahmedin Jemal. Cancer statistics, 2012. *CA Cancer J Clin*, 62(1):10–29, 2012.

[16] Soonmyung Paik, Steven Shak, Gong Tang, Chungyeul Kim, Joffre Baker, Maureen Cronin, Frederick L. Baehner, Michael G. Walker, Drew Watson, Taesung Park, William Hiller, Edwin R. Fisher, D Lawrence Wickerham, John Bryant, and Norman Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*, 351(27):2817–2826, Dec 2004.

[17] S. Badve and H. Nakshatri. Oestrogen-receptor-positive breast cancer: towards bridging histopathological and molecular classifications. *J Clin Pathol*, 62(1):6–12, Jan 2009.

[18] C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, Nov 1991.

[19] H. J. Bloom and W. W. Richardson. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *Br J Cancer*, 11(3):359–377, Sep 1957.

[20] C. Genestie, B. Zafrani, B. Asselain, A. Fourquet, S. Rozan, P. Validire, A. Vincent-Salomon, and X. Sastre-Garau. Comparison of the prognostic value of scarff-bloom-richardson and nottingham histological grades in a series of 825 cases of breast cancer: major importance of the mitotic count as a component of both grading systems. *Anticancer Res*, 18(1B):571–576, 1998.

[21] P. Boiesen, P. O. Bendahl, L. Anagnostaki, H. Domanski, E. Holm, I. Idvall, S. Johansson, O. Ljungberg, A. Ringberg, G. Ostberg, and M. Fernö. Histologic

grading in breast cancer–reproducibility between seven pathologic departments. south sweden breast cancer group. *Acta Oncol*, 39(1):41–45, 2000.

[22] L. W. Dalton, S. E. Pinder, C. E. Elston, I. O. Ellis, D. L. Page, W. D. Dupont, and R. W. Blamey. Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Mod Pathol*, 13(7):730–735, Jul 2000.

[23] John S Meyer, Consuelo Alvarez, Clara Milikowski, Neal Olson, Irma Russo, Jose Russo, Andrew Glass, Barbara A Zehnbauer, Karen Lister, Reza Parwaresch, and Cooperative Breast Cancer Tissue Resource. Breast carcinoma malignancy grading by bloom-richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. *Mod Pathol*, 18(8):1067–1078, Aug 2005.

[24] A. J. M. Connor, S. E. Pinder, C. W. Elston, J. A. Bell, P. Wencyk, J. F. R. Robertson, R. W. Blarney, R. I. Nicholson, and I. O. Ellis". Intratumoural heterogeneity of proliferation in invasive breast carcinoma evaluated with mibi antibody. *The Breast*, 6(4):171 – 176, 1997.

[25] Lurdes Torres, Franclim R. Ribeiro, Nikos Pandis, Johan A. Andersen, Sverre Heim, and Manuel R. Teixeira. Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res Treat*, 102(2):143–155, Apr 2007.

[26] Marco Gerlinger, Andrew J. Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q. McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R. Santos, Mahrokh Nohadani, Aron C. Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P Andrew Futreal, and Charles Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*, 366(10):883–892, Mar 2012.

[27] Yi-Hsuan Hsiao, Ming-Chih Chou, Carol Fowler, Jeffrey T Mason, and Yan-Gao Man. Breast cancer heterogeneity: mechanisms, proofs, and implications. *J Cancer*, 1:6–13, 2010.

[28] Marina V. Zavyalova, Evgeny V. Denisov, Lubov A. Tashireva, Tatiana S. Gerashchenko, Nikolay V. Litviakov, Nikolay A. Skryabin, Sergey V. Vtorushin, Nadezhda S. Telegina, Elena M. Slonimskaya, Nadezhda V. Cherdyntseva, and Vladimir M. Perelmuter. Phenotypic drift as a cause for intratumoral morphological heterogeneity of invasive ductal breast carcinoma not otherwise specified. *Biores Open Access*, 2(2):148–154, Apr 2013.

[29] J.F. O'Callaghan. An alternative definition for "neighborhood of a point". *IEEE Trans. on Computers*, 24(11):1121–1125, 1975.

[30] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *Proc. 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 496–499, 2008.

[31] B. Weyn, G. van de Wouwer, A. van Daele, P. Scheunders, D. van Dyck, E. van Marck, and W. Jacob. Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry*, 33(1):32–40, Sep 1998.

[32] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(6):610–621, 1973.

[33] R.E. Bellman and Rand Corporation. *Dynamic programming*. Rand Corporation research study. Princeton University Press, 1957.

[34] Hanchuan Peng, Fuhui Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, 2005.

[35] A. C. Evans, J. A. Frank, J. Antel, and D. H. Miller. The role of mri in clinical trials of multiple sclerosis: comparison of image processing techniques. *Ann Neurol*, 41(1):125–132, Jan 1997.

[36] D. S. Shin, N. B. Javornik, and J. W. Berger. Computer-assisted, interactive fundus image processing for macular drusen quantitation. *Ophthalmology*, 106(6):1119–1125, Jun 1999.

[37] Amit Vasanji and Brett A. Hoover. Art & science of imaging analytics. *Applied Clinical Trials*, 22(3):38, March 2013.

[38] Scott Doyle, James Monaco, Michael Feldman, John Tomaszewski, and Anant Madabhushi. An active learning based classification strategy for the minority class problem: application to histopathology annotation. *BMC Bioinformatics*, 12:424, 2011.

[39] Daniel Berrar, Ian Bradbury, and Werner Dubitzky. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics*, 22(10):1245–1250, 2006.

[40] L. Didaci, G. Giacinto, F. Roli, and G.L. Marcialis. A study on the performances of dynamic classifier selection based on local accuracy estimation. *Pattern Recognition*, 38, 2005.

[41] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, 2001.

[42] L. G. Nyúl, J. K. Udupa, and X. Zhang. New variants of a method of mri scale standardization. *IEEE Trans. Med. Imag.*, 19(2):143–150, Feb 2000.

[43] Anant Madabhushi and Jayaram K. Udupa. New methods of mr image intensity standardization via generalized scale. *Med Phys*, 33(9):3426–3434, Sep 2006.

[44] Daniel Palumbo, Brian Yee, Patrick O'Dea, Shane Leedy, Satish Viswanath, and Anant Madabhushi. Interplay between bias field correction, intensity standardization, and noise filtering for t2-weighted mri. *Conf Proc IEEE Eng Med Biol Soc*, 2011:5080–5083, 2011.

[45] Scott Doyle, Michael Feldman, John Tomaszewski, and Anant Madabhushi. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Trans. Biomed. Eng.*, 59(5):1205–1218, 2012.

[46] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013.

[47] Yukako Yagi. Color standardization and optimization in whole slide imaging. *Diag Pathol*, 6 S.1:S15, 2011.

[48] G. D. Finlayson, S. D. Hordley, and P. M. HubeL. Color by correlation: a simple, unifying framework for color constancy. *IEEE Trans. Pattern Anal. and Machine Intel.*, 23(11):1209–1221, 2001.

[49] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[50] Larry Latson, Bruce Sebek, and Kimerly A Powell. Automated cell nuclear segmentation in color images of hematoxylin and eosin-stained breast biopsy. *Anal Quant Cytol Histol*, 25(6):321–331, Dec 2003.

[51] Sokol Petushi, Fernando U Garcia, Marian M Haber, Constantine Katsinis, and Aydin Tozeren. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC Med Imaging*, 6:14, 2006.

[52] James P Monaco, John E Tomaszewski, Michael D Feldman, Ian Hagemann, Mehdi Moradi, Parvin Mousavi, Alexander Boag, Chris Davidson, Purang Abolmaesumi, and Anant Madabhushi. High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models. *Med Image Anal*, 14(4):617–629, Aug 2010.

[53] H. Fatakdawala, Jun Xu, A. Basavanhally, G. Bhanot, S. Ganesan, M. Feldman, J. E. Tomaszewski, and A. Madabhushi. Expectation–maximization-driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology. *IEEE Trans. Biomed. Eng.*, 57(7):1676–1689, 2010.

[54] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *Proc. 5th IEEE Int. Symp. Biomedical Imaging: From Nano to Macro ISBI 2008*, pages 284–287, 2008.

[55] Adel Hafiane, Filiz Bunyak, and Kannappan Palaniappan. Fuzzy clustering and active contours for histopathology image segmentation and nuclei detection. *Lect Notes Comput Sci*, 5259:903–914, Oct 2008.

[56] Jun Xu, Andrew Janowczyk, Sharat Chandran, and Anant Madabhushi. A weighted mean shift, normalized cuts initialized color gradient based geodesic active contour model: applications to histopathology image segmentation. In Benoit M. Dawant and David R. Haynor, editors, *SPIE Medical Imaging*, volume 7623 (1), page 76230Y. SPIE, 2010.

[57] Andrew Janowczyk, Sharat Chandran, Rajendra Singh, Dimitra Sasaroli, George Coukos, Michael D Feldman, and Anant Madabhushi. Hierarchical normalized cuts: unsupervised segmentation of vascular biomarkers from ovarian cancer tissue microarrays. *Med Image Comput Comput Assist Interv*, 12(Pt 1):230–238, 2009.

[58] K. Kayser, M. Shaver, F. Modlinger, K. Postl, and J. J. Moyers. Neighborhood analysis of low magnification structures (glands) in healthy, adenomatous, and carcinomatous colon mucosa. *Pathol Res Pract*, 181(2):153–158, May 1986.

[59] Nicolas Loménie and Daniel Racoceanu. Point set morphological filtering and semantic spatial configuration modeling: Application to microscopic image and bio-structure analysis. *Pattern Recogn.*, 45(8):2894–2911, August 2012.

[60] Cigdem Demir, S. Humayun Gultekin, and Bülent Yener. Augmented cell-graphs for automated cancer diagnosis. *Bioinformatics*, 21 Suppl 2:ii7–i12, Sep 2005.

[61] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, J.E. Tomaszewski, and M. Feldman. Automated grading of prostate cancer using architectural and textural image features. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1284–87, Washington DC, 2007.

[62] Chao-Hui Huang, Antoine Veillard, Ludovic Roux, Nicolas Loménie, and Daniel Racoceanu. Time-efficient sparse analysis of histopathological whole slide images. *Comput Med Imaging Graph*, 35(7-8):579–591, 2011.

[63] Sahirzeeshan Ali, Robert Veltri, Jonathan A Epstein, Christhunesa Christudass, and Anant Madabhushi. Cell cluster graph for prediction of biochemical recurrence in prostate cancer patients from tissue microarrays. In *SPIE Medical Imaging*, pages 86760H–86760H. International Society for Optics and Photonics, 2013.

[64] A. E. Dawson, RE Austin, Jr, and D. S. Weinberg. Nuclear grading of breast carcinoma by image analysis. classification by multivariate and neural network analysis. *Am J Clin Pathol*, 95(4 Suppl 1):S29–S37, Apr 1991.

[65] B. Palcic. Nuclear texture: can it be used as a surrogate endpoint biomarker? *J Cell Biochem Suppl*, 19:40–46, 1994.

[66] Wei Wang, John A. Ozolek, and Gustavo K. Rohde. Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry A*, 77(5):485–494, May 2010.

[67] S. S. Cross. Fractals in pathology. *J Pathol*, 182(1):1–8, May 1997.

[68] Pranab Dey and Sambit K. Mohanty. Fractal dimensions of breast lesions on cytology smears. *Diagn Cytopathol*, 29(2):85–86, Aug 2003.

[69] Mauro Tambasco, Misha Eliasziw, and Anthony M. Magliocco. Morphologic complexity of epithelial architecture for predicting invasive breast cancer survival. *J Transl Med*, 8:140, 2010.

[70] Chiang Hau Tay, Ramakrishnan Mukundan, and Daniel Racoceanu. Multifractal analysis of histopathological tissue images. In *Image and Vision Computing New Zealand*, 2011.

[71] George Lee, Carlos Rodriguez, and Anant Madabhushi. Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. *IEEE/ACM Trans Comput Biol Bioinform*, 5(3):368–384, 2008.

[72] R.E. Schapire. The boosting approach to machine learning: An overview. In *Nonlin. Est. and Class.*, 2002.

[73] Satish E. Viswanath, Nicholas B. Bloch, Jonathan C. Chappelow, Robert Toth, Neil M. Rofsky, Elizabeth M. Genega, Robert E. Lenkinski, and Anant Madabhushi. Central gland and peripheral zone prostate tumors have significantly different quantitative imaging signatures on 3 tesla endorectal, in vivo t2-weighted mr imagery. *J Magn Reson Imaging*, 36(1):213–224, Jul 2012.

[74] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185–205, Apr 2005.

[75] Zhisong He, Jian Zhang, Xiao-He Shi, Le-Le Hu, Xiangyin Kong, Yu-Dong Cai, and Kuo-Chen Chou. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, 5(3):e9603, 03 2010.

[76] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognit*, 42(6):1093–1103, Jun 2009.

[77] Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Robert J. Marinelli, Torsten O. Nielsen, Marc J. van de Vijver, Robert B. West, Matt van de Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*, 3(108):108–113, Nov 2011.

[78] Mauro Tambasco and Anthony M. Magliocco. Relationship between tumor grade and computed architectural complexity in breast cancer specimens. *Hum Pathol*, 39(5):740–746, May 2008.

[79] Jun Kong, O. Sertel, H. Shimada, K. Boyer, J. Saltz, and M. Gurcan. Computer-aided grading of neuroblastic differentiation: Multi-resolution and multi-classifier approach. In *Proc. IEEE Int. Conf. Image Processing ICIP 2007*, volume 5, 2007.

[80] Giuseppe Boccignone, Paolo Napoletano, Vittorio Caggiano, and Mario Ferraro. A multiresolution diffused expectation-maximization algorithm for medical image segmentation. *Computers in Biology and Medicine*, 37(1):83 – 96, 2007.

[81] CJ Adcock. Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2):261–283, 1997.

[82] Sayan Mukherjee, Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R. Golub, and Jill P. Mesirov. Estimating dataset size requirements for classifying dna microarray data. *J Comput Biol*, 10(2):119–142, 2003.

[83] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002.

[84] Frank J. Brooks and Perry W. Grigsby. Quantification of heterogeneity observed in medical images. *BMC Med Imaging*, 13:7, 2013.

[85] Ulrika Wickenberg-Bolin, Hanna Göransson, Mårten Fryknäs, Mats G Gustafsson, and Anders Isaksson. Improved variance estimation of classification performance via reduction of bias caused by small sample size. *BMC Bioinformatics*, 7(1):127, 2006.

[86] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2-3):133–168, 1997.

[87] Austin J. Ramme, Nicole DeVries, Nicole A. Kallemyn, Vincent A. Magnotta, and Nicole M. Grosland. Semi-automated phalanx bone segmentation using the expectation maximization algorithm. *J Digit Imaging*, 22(5):483–491, Oct 2009.

[88] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):pp. 80–83, 1945.

[89] Hervé Abdi. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, 2007.

[90] A. C. Ruifrok and D. A. Johnston. Quantification of histochemical staining by color deconvolution. *Anal Quant Cytol Histol*, 23(4):291–299, Aug 2001.

[91] Jun Xu, Andrew Janowczyk, Sharat Chandran, and Anant Madabhushi. A high-throughput active contour scheme for segmentation of histopathological imagery. *Medical image analysis*, 15(6):851–862, 2011.

[92] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[93] Laurent D. Cohen. On active contour models and balloons. *CVGIP: Image Underst.*, 53(2):211–218, 1991.

[94] C. Li, C. Xu, C. Gui, and M. D. Fox. Distance regularized level set evolution and its application to image segmentation. *Image Processing, IEEE Transactions on*, 19(12):3243 –3254, December 2010.

[95] Jun Xu, Rachel Sparks, Andrew Janowczyk, John E Tomaszewski, Michael D Feldman, and Anant Madabhushi. High-throughput prostate cancer gland detection, segmentation, and classification from digitized needle core biopsies. In *Prostate Cancer Imaging. Computer-Aided Diagnosis, Prognosis, and Intervention*, pages 77–88. Springer, 2010.

[96] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[97] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[98] A. Fitzgibbon, M. Pilu, and R. B. Fisher. Direct least square fitting of ellipses. *IEEE Trans on Patt Anal and Machine Intel*, 21(5):476–480, 1999.

[99] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[100] J. Sudbø, A. Bankfalvi, M. Bryne, R. Marcelpoil, M. Boysen, J. Piffko, J. Hemmer, K. Kraft, and A. Reith. Prognostic value of graph theory-based tissue architecture analysis in carcinomas of the tongue. *Lab Invest*, 80(12):1881–1889, Dec 2000.

[101] A. K. Jain. *Fundamentals of Digital Image Processing.* Prentice Hall, 1989.

[102] N. Weidner, J. P. Semple, W. R. Welch, and J. Folkman. Tumor angiogenesis and metastasis–correlation in invasive breast carcinoma. *N Engl J Med*, 324(1):1–8, Jan 1991.

[103] Aissar Eduardo Nassif and Renato Tâmbara Filho. Immunohistochemistry expression of tumor markers cd34 and p27 as a prognostic factor of clinically localized prostate adenocarcinoma after radical prostatectomy. *Rev Col Bras Cir*, 37(5):338–344, Oct 2010.

[104] Boban M. Erovic, Csilla Neuchrist, Uwe Berger, Karem El-Rabadi, and Martin Burian. Quantitation of microvessel density in squamous cell carcinoma of the head and neck by computer-aided image analysis. *Wien Klin Wochenschr*, 117(1-2):53–57, Jan 2005.

[105] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[106] J. Scheidler, H. Hricak, D. B. Vigneron, K. K. Yu, D. L. Sokolov, L. R. Huang, C. J. Zaloudek, S. J. Nelson, P. R. Carroll, and J. Kurhanewicz. Prostate cancer: localization with three-dimensional proton mr spectroscopic imaging–clinicopathologic study. *Radiology*, 213(2):473–480, Nov 1999.

[107] P. Tiwari, S. Viswanath, J. Kurhanewicz, A. Sridhar, and A. Madabhushi. Multimodal wavelet embedding representation for data combination (maweric): integrating magnetic resonance imaging and spectroscopy for prostate cancer detection. *NMR Biomed*, 25(4):607–619, Apr 2012.

[108] John Kurhanewicz, Mark G. Swanson, Sarah J. Nelson, and Daniel B. Vigneron. Combined magnetic resonance imaging and spectroscopic imaging approach to molecular imaging of prostate cancer. *J Magn Reson Imaging*, 16(4):451–463, Oct 2002.

[109] Leo Breiman. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383, 1996.

[110] Gang Wu and Edward Y Chang. Class-boundary alignment for imbalanced dataset learning. In *ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC*, pages 49–56, 2003.

[111] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[112] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.