DEVELOPING IMAGE INFORMATICS METHODS FOR HISTOPATHOLOGICAL COMPUTER-AIDED DECISION SUPPORT SYSTEMS

A Dissertation
Presented to
The Academic Faculty

by

Sonal Kothari

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December, 2013

DEVELOPING IMAGE INFORMATICS METHODS FOR HISTOPATHOLOGICAL COMPUTER-AIDED DECISION SUPPORT SYSTEMS

Approved by:

Dr. May D. Wang, Advisor
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Patricio A. Vela, Co-Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Anthony J. Yezzi
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Paul J. Benkeser
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Andrew F. Peterson
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Andrew N. Young
Pathology & Laboratory Medicine
Emory University

Date Approved: October 04, 2013

To my loving parents and fiancé

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. May D. Wang, for giving me the opportunity to work in her lab, to support my research, and to constantly challenge my intellectual capabilities. I admire her vision about high-impact research, her strength to work endlessly and her ambition to be the best in every project our lab pursues. Although last few years have been the busiest and most hard working years of my life so far, I am really grateful of everything I have accomplished under her guidance.

I would like to thank all Bio-MIBLab members—graduate students, undergraduate students and post-docs—for establishing a stimulating and diverse research environment. I am grateful for all the feedback during my lab presentations. I would like to especially thank my research mentors—Dr. Qaiser Chaudry and Dr. John Phan—for their immense patience to answer my questions and for their guidance throughout the research process. I would like to thank other students in the lab—Quo, Chanchala, James, Leo, and Janani—for being there and making me laugh even on the toughest days.

I would like to thank my committee members for the valuable feedback and for taking time out of their busy schedule to examine and critique my research. I would like to especially thank Dr. Andrew Young for his insight and motivation throughout my research in last five years. I am grateful for his guidance.

Finally, I would like thank my family for always being supportive. I would like to thank my role model in life—my dad, Dr Gautam Chand Kothari—for really understanding me and always being so patient with me. I am grateful to him for motivating me to work towards the target without being distracted by small problems in life. I would like to thank my mom, Pushpa Kothari, for always listening to me and reminding me that I am doing my best, so I shouldn't worry about the consequences. I would like to thank my fiancé, John, for providing me confidence and making me smile every day. I would like to thank my sisters (Priya and Priyanka), brother-in-laws (Dhananjay and Ramit), and nieces (Swara, Anushka, and Reet) for being a constant source of happiness in my life. I would like thank my best friend and cousin, Neha, for believing I can achieve anything.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS AND ABBREVIATIONS

ARI	Adjusted Rand index
CC	Clear Cell
CH	Chromophobe
CDSS	Clinical decision support system
Clust	Clustering-based tissue-fold detection
ConnSoftT	Connectivity-based soft threshold
CV	Cross-validation
GLCM	Gray-level Co-occurrence Matrix
Gbm	Glioblastoma
H&E	Hematoxylin and eosin
LDA	Linear discriminant analysis
ON	Oncocytoma
OvCa	Ovarian serous adenocarcinoma
k-NN	k nearest neighbors
KiCa	Kidney clear cell carcinoma
mRMR	Minimum redundancy maximum relevance
MultiQ	Multivariate quantization
MultiSubQ	Multivariate subset quantization
PA	Papillary
PatientViz	Patient-level visualization
PCA	Principal component analysis
PVCA	Principal variation component analysis

RCC	Renal cell carcinoma
ROI	Region-of-interest
SoftT	Soft threshold
SOM	Self-organizing map
SVM	Support Vector Machines
TCGA	The Cancer Genome Atlas
TissueViz	Tissue-level visualization
TNR	True negative rate
TPR	True positive rate
UniQ	Univariate quantization
WSI	Whole-slide image

SUMMARY

This dissertation focuses on developing imaging informatics algorithms for clinical decision support systems (CDSSs) based on histopathological whole-slide images (WSIs). Currently, histopathological analysis is a common clinical procedure for diagnosing cancer presence, type, and progression. While diagnosing patients using biopsy slides, pathologists manually select the most progressed cancer regions and assess nuclear morphology. However, making decisions manually from a slide with millions of nuclei can be time-consuming and subjective. Researchers have proposed CDSSs that help in decision making by quantifying morphological properties in portions of biopsy slides selected by a pathologist. However, existing CDSSs have not been widely used in clinical practice because of the following limitations: (1) Human intervention is required for region-of-interest (ROI) selection and CDSS operation and (2) The performance of CDSSs is not robust and is sensitive to data variance [1, 2]. The development of robust CDSSs for WSIs, without ROI selection, faces several informatics challenges: (1) Lack of robust segmentation methods for histopathological images, (2) Semantic gap between quantitative information and pathologist's knowledge, (3) Lack of batch-invariant imaging informatics methods, (4) Lack of knowledge models for capturing informative patterns in large WSIs, and (5) Lack of guidelines for optimizing and validating diagnostic models.

Recently, a large collection of WSIs with linked genomic and clinical data was provided by a public repository—The Cancer Genome Atlas (TCGA)—to facilitate in-depth understanding and treatment of cancer [3]. TCGA enables data-driven research. I

conducted advanced imaging informatics research to extract information from WSIs, to model knowledge embedded in these large datasets, and to assist decision making with biological and clinical validation.

Downstream knowledge modeling for cancer diagnosis requires that upstream information extraction methods are reproducible, invariant (to acquisition artifacts and batch effects), and comprehensive. I developed the following robust segmentation techniques: (1) A connectivity-based threshold model, which adapts with tissue variation, for tissue-fold segmentation; (2) A supervised stain segmentation system, which normalizes and segments test image using multiple reference images; and (3) An edge-based nuclear cluster segmentation method, which segments nuclei using concavity detection and ellipse fitting. In addition, I developed a comprehensive image feature set that represents many aspects of histopathological images including color, shape, texture, and topology. Finally, I developed image-level and information-level normalization methods to address various forms of batch-effects.

Modeling knowledge in large WSIs is hindered by the biological variation within the images and the semantic gap between quantitative information and pathologists' knowledge. I developed a tissue-level visualization tool, called TissueViz, which facilitates the study of spatial patterns and the identification of ROI in WSIs through three visualization modes: single feature variation, unsupervised multi-dimensional clustering, and supervised classification. Thereafter, I modeled the knowledge about the

spatial structure in WSIs using quantization methods, which quantize image feature space and quantify percent of WSIs in different quantization blocks.

The novel knowledge models generated by my imaging informatics algorithms enable decision making. I validated the knowledge models for two applications: (1) diagnosis of histopathology-based endpoints such as subtype and grade and (2) prediction of clinical endpoints such as metastasis, stage, lymphnode spread, and survival. The statistically emergent feature subsets in the models for histopathology-based endpoints complied with pathologists' knowledge. For example, nuclear shape features were overrepresented in Fuhrman nuclear grade model for kidney carcinoma. Besides validating models based on histopathological features, I also validated models for prognosis prediction using histopathological and genomic features. Using these predictive models, I found genomic and imaging markers for patient stratification. I developed a patient-level visualization tool, called PatientViz, which empowers the discovery of separable, reproducible, and prognostically significant clusters among cancer patients using histopathological knowledge.

From this research, I have attempted to provide evidence that pathology imaging informatics algorithms assist and enhance clinical decision making in cancer. I have illustrated results for ovarian and kidney carcinoma but all methods can be easily extended to other types of cancer. Automatic, batch-invariant, and comprehensive imaging informatics algorithms validated by biological interpretation of cancer endpoints can provide a deeper understanding of cancer histopathology.

CHAPTER 1

INTRODUCTION

The primary goal of this dissertation is to develop imaging informatics methods for CDSSs based on histopathological WSIs. This chapter describes the field of pathology image informatics and discusses state-of-art methods for CDSSs. Most of the content and figures in this chapter is a part of a review article on pathology imaging informatics [4].

To establish the motivation of this dissertation, this chapter highlights the challenges posed by histopathology and discusses the importance of imaging informatics in pathology. Next, to establish the potential impact of this dissertation, this chapter discuss current challenges and state-of-art methods for the various blocks of a pathology imaging informatics pipeline including (1) quality control of histopathological images, (2) information extraction that captures image properties at the pixel-, object-, and semantic-levels, (3) data and information visualization that explores WSIs for de novo discovery, and (4) knowledge modeling that utilizes image features to model meaningful knowledge for decision making, and (4) decision making that develops prediction models for diagnostic or prognostic applications (Figure 1). Thereafter, the chapter discusses state-of-art commercial systems for whole-slide image analysis and their limitations. Finally, the chapter describes the specific aims and structure of this dissertation.

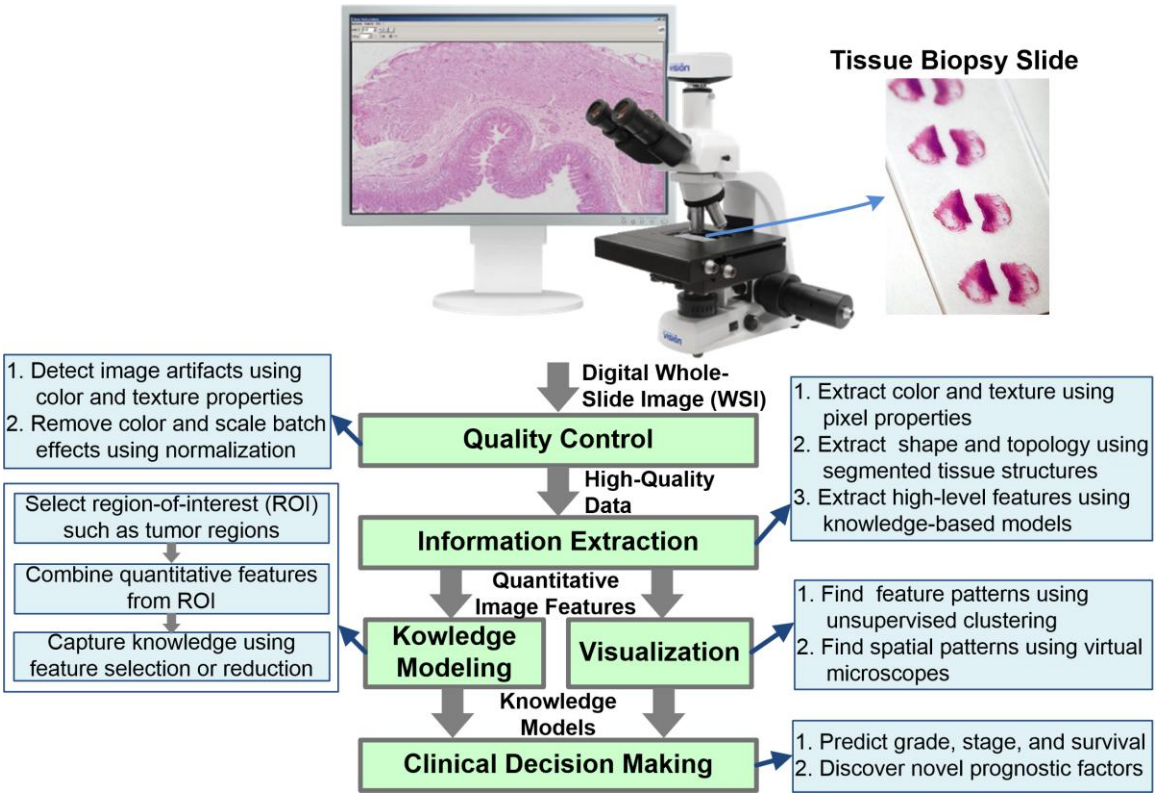

Figure 1: Translational pathology informatics pipeline for analysis of WSIs.
This pipeline has the following key components: quality control to insure only high-quality data is processed, information extraction to covert WSIs into quantitative features, visualization to interpret the image feature space and find patterns, knowledge modeling to model knowledge in WSIs, and decision making to develop clinical diagnostic and prognostic models.

Histopathological Images

Histopathology is the study of microscopic anatomical changes in diseased tissue samples. Tissue samples are usually obtained during surgery, biopsy or autopsy. Then they are fixed using either paraffin embedding or cryostat freezing and sectioned into very thin slices. These slices are then placed on a glass slide and stained with one or more stains. The goal of staining is to highlight cellular structures for study using light microscopy. Hematoxylin and Eosin (H&E) staining protocol is the most commonly used protocol for morphological analysis of tissue samples. H&E staining enhances four colors in histopathological images: blue-purple, white, pink and red (Figure 2). These colors correspond to specific cellular structures. Basophilic structures containing nucleic acids—ribosome and nuclei—tend to stain blue-purple; eosinphilic intra- and extracellular proteins in cytoplasmic regions tend to stain bright pink; empty spaces—the lumen of glands—do not stain and tend to be white; and red blood cells stain intensely red.

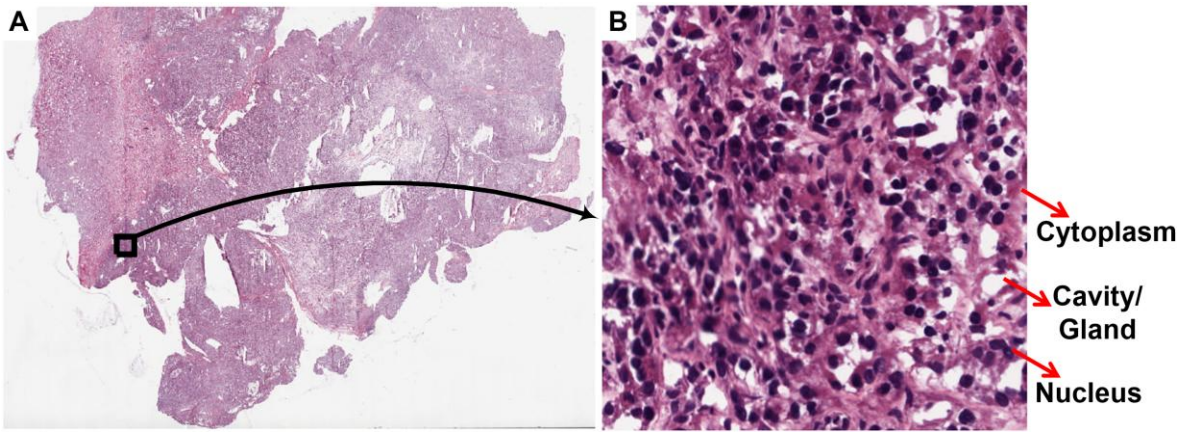

Figure 2: A sample histopathological image stained with H&E stains.
(A) WSI, and (B) 512x512-pixel rectangular section, where nuclei, cytoplasm, and glands appear blue-purple, pink, and white, respectively.

Histopathological analysis of tissue samples has several clinical applications—1) confirmation of a disease, 2) exclusion of a disease, 3) assessment of subtype and extent of a disease. Specifically, while analyzing a tumor tissue sample, pathologists confirm the presence of cancer, its subtype, its grade, and other cancer-specific prognostic indicators such mitotic counts [5, 6].

Importance of Pathology Imaging Informatics

Pathology imaging informatics refers to the analytical and computational methods for handling, analyzing, and exploring histopathological images and their associated clinical data [5, 7-11]. Imaging informatics play the following two important roles in the field of pathology. Firstly, it facilitates the development of CDSSs for patient diagnosis. Secondly, it aids in discovery novel biomarkers for research applications.

CDSSs provide a fast, objective, and reproducible means for histopathological analysis. Most of the existing CDSSs focus on images that represent only portions of tissue slides rather than on whole-slide images (WSIs) [5]. In comparison to WSIs, typical histopathological images reflect portions of tissue slides that are pathologist-selected portions with higher quality, and represent the most informative and disease-relevant regions of tissue slides [12]. Therefore, CDSSs based on tissue slide portions face limited challenges in the areas of quality control, ROI selection, and computational complexity. However, images of slide portions do not capture the complete information available to the pathologist during initial microscopic analysis. Moreover, they are subject to biases related to the knowledge of the pathologist that selected the image portions [12]. Thus, this dissertation focuses on methods for analysis of WSIs.

Patient-level prediction modeling and exploratory analysis is important for a number of clinical applications including diagnostics and therapeutics [13]. The importance of accurate image-based disease diagnosis and the development of novel pathology informatics techniques have led to the establishment of databases such as the NCI Cooperative Prostate Cancer Tissue Resource (CPCTR) [14], TCGA [3], and the Human Protein Atlas (HPA) [15]. Such databases provide a large number of high-quality histopathological images and associated clinical data, further stimulating the development of novel informatics methods. Some of these databases also provide matched genomic and proteomic data, enabling multi-modal studies that associate "–omic" data with histopathological image features. To use the full potential of these repositories, it is essential to develop robust pathology imaging informatics methods.

Quality Control

The quality of histopathological images is usually affected by (1) artifacts acquired during image acquisition and (2) batch effects resulting from variations in experimental protocol. Both of these quality issues can affect the results of downstream clinical applications. Data quality is especially challenging in collaborative repositories, such as TCGA, where a large amount of high-throughput data is collected at multiple institutions [3]. To use the full potential of these repositories, it is essential that researchers develop robust quality correction methods for WSIs. This dissertation discusses the causes and effects of these quality issues and describe some existing methods for identifying, eliminating, and correcting them.

Image Artifacts

Errors in biopsy-slide preparation or microscope parameters may lead to anomalies, known as image artifacts, in WSIs. Common image artifacts include tissue folds, blurred regions, pen marks, shadows, and chromatic aberrations [11, 16]. Image artifacts have unpredictable effects on image segmentation and other quantitative image features. Therefore, it is essential to either eliminate or correct these artifacts. Tissue-fold artifacts, caused by layering of non-adherent tissue on the slide, can be eliminated using methods based on color-saturation and intensity [17, 18]. Figure 2 illustrates tissue folds and pen marks in WSIs. Blurred regions, caused by loss of microscope focus, can be eliminated using methods that detect sharpness or texture properties [19]. Chromatic aberrations occur when light dispersion through the microscopic lens varies with colors, leading to ghost colors along the edges of objects or discontinuities in an image. Wu et al. suggest a method that quantifies the amount of color dispersion at the object edges and realigns color components to correct chromatic aberration [20].

Although artifact correction/elimination is essential for robust downstream analysis, literature on the topic is relatively sparse. Moreover, most proposed methods have only been tested on a limited set of images as a proof-of-concept. Therefore, this dissertation presents a novel tissue-fold artifact detection method and validates it on a large set of WSIs.

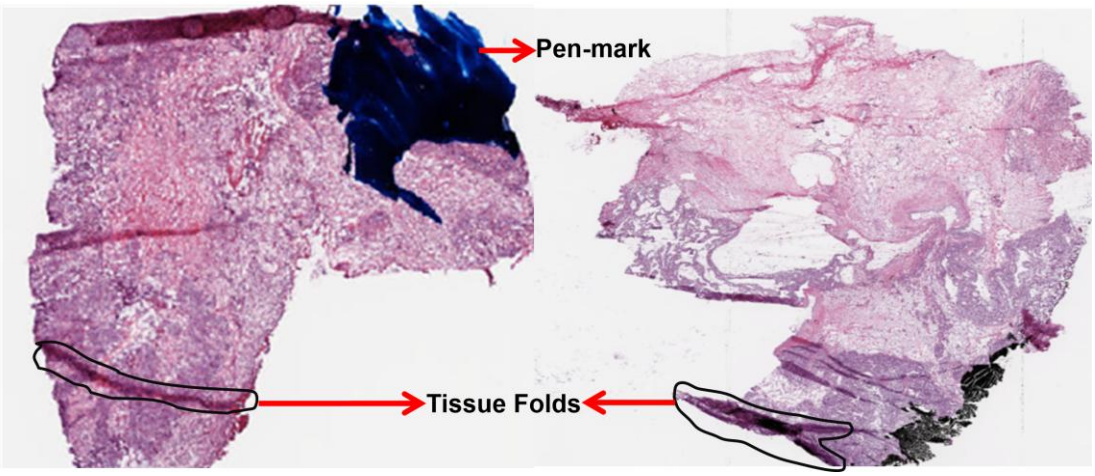

Figure 3: Tissue fold and pen-mark artifacts in WSIs.

Batch Effects

Differences in slide preparation, microscope, and digitizing device between two batches of data may lead to differences in image properties between the two batches. These differences, called batch effects, can bias the performance estimates of predictive models. Histopathological images often suffer from color and scale batch effects. Color batch effects can be addressed by normalizing the color of an image to a reference image [21, 22] or by converting the image to a color space that is not affected by color batch effects [23-25]. Unlike color batch effects, which affect only color properties of an image, scale batch effects can affect a variety of image features such as object size, topology, and texture. However, scale batch effects may be difficult to detect or to correct because apparent changes in scale may be induced by biological factors such as cancer grade or subtype.

Studies suggest that batch effects, if left un-corrected, can severely reduce the performance of genomic prediction models [26, 27]. Even though preliminary investigations, discussed above, suggest that batch effects are present in histopathological images as well, most researchers validate their diagnostic models on a single image dataset collected during a single experimental set-up. This dissertation presents novel methods for batch-effect normalization and discusses cross-batch performance with and without normalization.

Information Extraction

Pathological images are described using following three types of information: (1) image acquisition metadata including imaging modality parameters, (2) clinical data including patient history, their disease and treatment, and (3) content-based image features. Content-based features may be very informative for quantitative prediction modeling and for exploratory analysis. Content-based features are categorized into three levels—pixel-, object-, and semantic-level features—based on the amount of raw data captured by the features and the biological interpretability of the features (Figure 4) [28, 29].

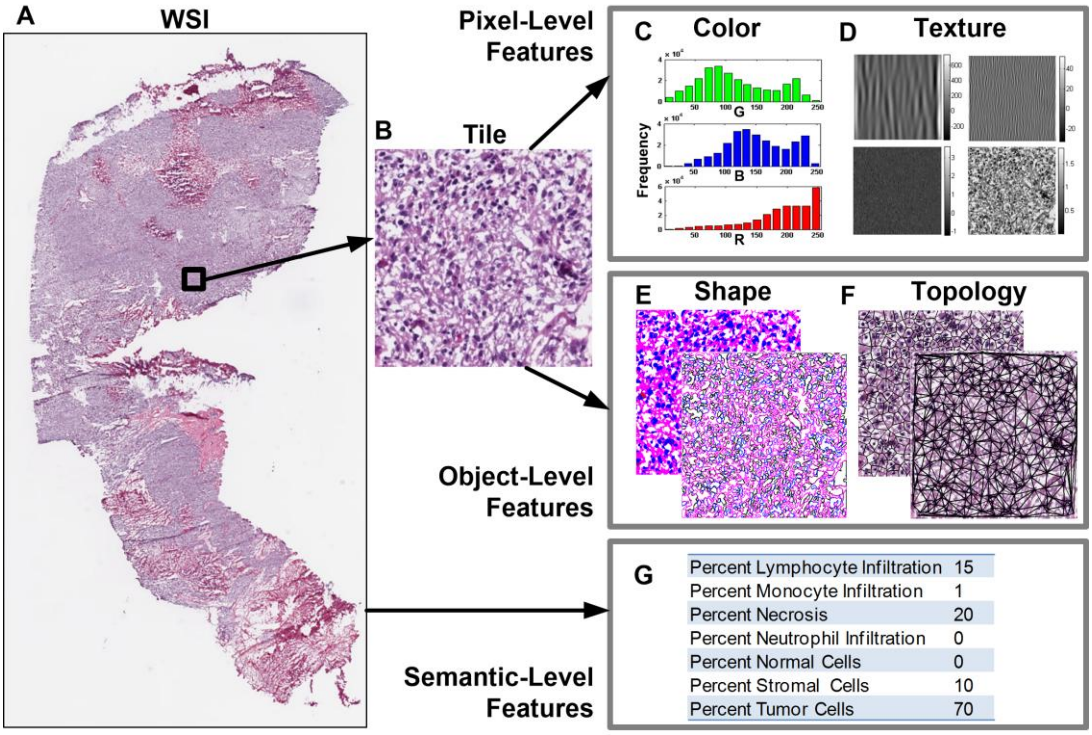

Figure 4: Overview of pixel-, object-, and semantic-level description of a histopathological WSI.
Representation of a WSI of a kidney renal clear cell carcinoma biopsy using various quantitative features extracted from a single image tile (B): pixel-level, including color histogram (C) and Gabor filter response (D); object-level, including segmented shapes (E) and graph-based topology (F); and semantic-level, including percentage of high-level clinical properties (G).

Pixel-Level Features

Pixel-level image features are in the lowest level of the information hierarchy because they are the least interpretable in terms of biology. Pixel-level image features do not focus on any specific set of pixels in a WSI. Rather, they consider all image pixels and capture properties such as color and texture. Color features quantify color spread, prominence and co-occurrence using statistics and frequencies of color histograms in different color spaces including RGB [30, 31], HSV [32], Luv [33], and Lab [24, 34] (Figure 4.C). As an example of the utility of color features, Celebi et al. used color features from eight color spaces to classify skin melanomas[35]. Texture features quantify image sharpness, contrast, changes in intensity, and discontinuities or edges by measuring properties derived from gray-level intensity profiles , Haralick Gray-level Co-occurrence Matrix (GLCM) features [25, 36, 37], wavelet sub-matrices [36, 38], multi-wavelet sub-matrices [37, 38], Gabor filter responses [25, 37] (Figure 4.D), and Fractals [37].Texture properties are generally extracted from grayscale images. However, Sertel et al. introduced the concept of color texture analysis by quantizing the colors in histopathological images using self-organizing maps (SOM) [39]. Numerous measures have been developed to capture image texture and each measure can include numerous parameters. This enables precise tuning of image features for various image processing applications. Unfortunately, most texture features are very difficult to interpret biologically. Thus, they are seldom used in knowledge-based models. Despite the lack of biological interpretability, pixel-level features are used extensively in data-driven models because they are simple to extract and are useful (at times sufficient) to describe the images. Figure 4 illustrates some pixel-level features of a kidney renal clear cell

carcinoma WSI, including RGB color histograms and Gabor filter textures at various scales. Pixel-level descriptors are computationally inexpensive to extract and are known to be diagnostically useful. Thus, researchers working on WSIs can easily adopt these features for image representation.

Object-Level Features

Object-level features are in a higher level of the information hierarchy compared to pixel-level features because they describe properties of the cellular structures—e.g., nuclei, cytoplasm, leukocytes, red blood cells, and glands—in a WSI. To extract object-based features, it is essential to first segment cellular structures. Since cellular structures appear in different colors in a stained histopathological sample, researchers have proposed color-based methods for segmentation. Literature supports both semi-automatic methods, with some user-interaction [36, 40], as well as completely automatic methods [41-43] for color segmentation of histopathological images. Most methods segment a pixel independent of its neighborhood. However, recently proposed methods consider pixel neighborhood properties using graph-cut [39], object-graph[44], and Markov models [45]. Color segmentation methods segment the nuclear stain in the image but nuclei often tend to overlap each other forming dense nuclear clusters. Therefore, another segmentation step is required to extract the nuclear objects. Previous work suggests edge-based [23, 46], region-based [47], and gradient-based [48] methods for nuclear cluster segmentation. The accuracy of image segmentation methods greatly impacts the robustness of quantitative image descriptors and downstream analysis. Figure 4.E

illustrates a pseudo colored segmentation mask, where blue, pink, and white represent nuclear, cytoplasmic, and no-stain/gland regions, respectively.

Object-level features describe the shape, texture, and spatial distribution of cellular structures in a WSI. Shape-based features can be broadly categorized into contour- and region-based features(Figure 4.E) [49]. Contour-based features include the properties of shape boundary such as perimeter, boundary fractal dimension, and bending energy. They also include coefficients or parameters of parametric shape models such Fourier shape descriptors and elliptical models. Region-based features include area, solidity, convex hull, Euler number, and Zernike moments [50]. Among all shape features, properties of elliptical-shape models of a nuclear boundary are most prevalent in pathology informatics because they are simple to extract and interpret and they have proven to be informative for predicting various cancer properties [39, 50-52].

Object-level texture features are similar to pixel-level texture features, except that they capture the texture of only a subset of image pixels associated with a tissue object. Nuclear texture is reported to be very informative for separating malignant regions [52], subtyping cancer [51], and grading cancer .

The spatial distribution of cellular structures, especially nuclei, in a tissue sample can be captured by topological or architectural features. A common technique for extracting topological properties involves development and characterization of spatial graphs, where graph nodes are centers of tissue objects. Researchers have found spatial graphs (e.g., Deluanay triangulations, Voronoi diagrams, minimum spanning trees,

Gabriel graphs, and Ulam trees) to be useful for extracting topological features in histopathological images (Figure 4.F). Common topological features include properties of spatial graphs (e.g., edge length, connectedness, and compactness). Besides graph-based properties, topological properties include object density, average distance between neighbors, and number of objects within a given neighborhood. Architectural features are useful for cancer endpoints such grading , classifying tumor vs. non-tumor regions [53, 54], classifying low vs. high lymphocytic infiltration regions [55], and predicting patient prognosis [56, 57].

Semantic-Level Features

Pixel- and object-level features capture useful information about histopathological images. However, they may be difficult to interpret biologically and are susceptible to noise. In contrast, semantic-level features capture easily interpretable high-level concepts such as presence or absence of nucleoli, necrosis, red blood cells, and leukocytes (Figure 4.G). Systems with semantic features can produce synoptic reports that detail high-level sample properties along with the final diagnosis. Therefore, if the final diagnosis appears questionable, it is easy to visually validate these properties on the image sample. A semantic feature is usually a classification or statistical rule based on a subset of low-level features (e.g., low-level properties such as nuclear texture, color, and gray level distribution may capture the high-level concept of nucleolus presence in a nucleus). Using these low-level features and some annotated data (in this case, annotated nucleus), a system could derive a classification rule for predicting the presence or absence of the high-level concept. Because not all low-level features may be useful for capturing high-level biological concepts, the system could use feature-preprocessing methods to select a

subset of the original or transformed features. Among these feature preprocessing methods, the bag-of-features method is the most commonly used for semantic features [58-60]. Bag-of-features analysis is especially common for SIFT (scale invariant feature transform) features and texton histograms. Development of an informatics system with semantic-level features requires a large amount of annotated training data. Thus, only a few image retrieval systems use semantic-level features [61-63].

Therefore, previous work suggests several useful image properties and descriptors for histopathological CDSSs However, there are certain limitations of the existing work—1) most automatic segmentation methods for segmenting cellular structures fail with the variations in tissue samples due to batch effects, 2) the semantic-level features are difficult to implement while most of the other existing features are difficult to interpret, and 3) most existing features are validated for only certain cancer endpoints and it is difficult to predict their performance for other endpoints. This dissertation presents a comprehensive set of image features including pixel- and object-level features. This set, which is composed of several high-level biologically interpretable subsets, is validated for several cancer endpoints.

Visualization

Visualization is an effective visual representation of data that can aptly communicate the information and aid learning. In the field of pathology informatics, visualizations are usually used to verify and interpret informative image features. High-dimensional data is very difficult and requires tools for interpreting the biological relevance of features and quantitative models. Large-scale studies such as TCGA aim to

reveal new insights about aggressive cancer endpoints and to discover new prognostically different subtypes.

Spatial Patterns

Images being spatial data, a prominent form of visualization is to display information directly on different regions of the image. With the availability of large histopathological data repositories such as TCGA, "virtual microscope" software applications have emerged that enable spatial exploration of high resolution digital WSIs [7, 64-66]. Without such applications, it is a challenge to share or even to view these images in real time. In addition, researchers have developed compression methods specifically for WSIs [67, 68]. The popularity of the Google Maps interface for exploring satellite images at many different detail levels has inspired similar tools for exploring whole slide tissue images [69-71]. In addition to viewing a WSI, some systems can highlight the regions-of-interest (e.g., regions of high-grade cancer or regions with lymphocyte infiltration) [45, 57, 72-75]. Moreover, some visualizations tag histopathological images with semantic labels such as necrosis, glands, and lymphocytes [62, 63] or highlight the spatial distributions of proteins, image features, or biomarker expression across the histopathological image [76].

High-Dimensional Feature Patterns

Patterns in image features can be captured in simple 2D or 3D visualizations such as scatter plots, distribution curves, box plots, histograms, and surface plots[36, 39, 52, 55, 57, 76].However, if the number of descriptors is very large (>50), such visualizations

may be difficult to implement or interpret. Thus, unsupervised clustering methods are often used to reduce feature space prior to visualization. Common clustering techniques in pathology imaging informatics include hierarchal clustering, SOMs, k-means, and expectation-maximization. Hierarchal clustering is useful for patient stratification and visualization [51, 65, 77-79]. SOMs, a neural network-based unsupervised learning method that reduces the high dimensional data to a low dimensional quantized form, is commonly used for feature interpretation [80], patient stratification [81] and segmentation [39, 82, 83] in pathology imaging informatics systems. The advantage of SOMs is that it provides a planar representation of its nodes, where the neighboring nodes are similar to each other. K-means, an optimization method that minimizes the sum of distances of samples to the center of the closest cluster, is mostly used for color segmentation [84] and bag-of-features representation of histopathological images. Bag-of-features representation of image samples is useful for image classification and visualization [59, 85]. Expectation-maximization, an iterative method that finds the maximum likelihood parameters for a model (mostly a mixture of Gaussians), is useful for segmenting histopathological images [24].

Both spatial- and patient-level visualizations of WSIs is an open area of research that requires interdisciplinary collaborations among pathologists, biologists, and computer scientists. Such collaboration is necessary to tackle the difficult problem of discovering and interpreting novel patterns in histopathological data that may lead to improved patient care. Moreover, it is necessary to develop novel quantitative metrics for

assessing the stability and reproducibility of patterns related to both spatial- and patient-level analysis to ensure that these patterns are biologically relevant.

Knowledge Modeling

After quality control and information extraction, the next step in WSI analysis is knowledge modeling. Knowledge modeling transforms information into meaningful and useful knowledge for decision making.

Here, three important steps of WSI prediction modeling are discussed: (1) region-of-interest selection and tile-based WSI representation, (2) informative feature selection and reduction, and (3) classification.

Region-of-Interest Selection

A high-resolution scan of a tissue biopsy slide results in a very large WSI (e.g., up to 40,000x60,000 pixels). Such WSIs contain a large amount of biologically related spatial variation including regions of high-grade tumor, low-grade tumor, necrosis, stroma, and lymphocyte-infiltrated tumor. When pathologists examine a WSI, they identify regions that are most important or relevant for the final prognostic decision (e.g., the region with the highest cancer grade). Similarly, an informatics system aims to identify a ROI in the WSI before developing a predictive model. Several researchers have developed supervised models for identifying ROIs in WSIs, but these methods require prior annotation for training [72, 73, 86]. Recently, researchers have proposed unsupervised knowledge-based methods for identifying ROIs [75, 87].

Because of limitations in computer memory and processing time, WSIs are often cropped into smaller tiles (e.g., 512x512-pixel tiles), and then features are extracted from each tile in parallel [24, 65, 73, 75, 88, 89]. After identifying tiles corresponding to ROIs, an informatics system can either combine the tiles to represent the WSI in a prediction model [51] or predict the label for individual tiles and then combine labels to represent the final prediction result of the WSI [24]. In the former method, outlier features might dominate WSI properties. In the latter method, annotation of individual tiles, instead of the WSI, might be necessary for training models. A related topic to piecewise analysis of WSIs is multi-resolution or multi-scale analysis, where a WSI is processed at various scales/resolutions to achieve different modeling objectives [24, 25, 74, 75]. The basic concept of multi-scale analysis is that a coarse level of prediction—such as tumor and non-tumor classification—can be achieved at a low resolution, where WSIs are smaller and processing time is shorter. In contrast, for more complex problems such as grade and subtype prediction, WSIs need to be processed at higher resolution. Representation of WSIs by combining data from multiple WSI tiles is an emerging area of research with limited published results, especially in the context of clinical prediction.

Most WSIs are millions of pixels in size and capture a large amount of biological heterogeneity. Thus, it is necessary to develop automatic methods for accurately selecting ROIs in WSIs. Without accurate ROI selection, prediction performance of CDSSs for WSIs may suffer compared to that for manually selected image portions. Besides ROI selection, innovative methods are needed for WSI representation that can capture the biological heterogeneity in patient samples. Such representation methods will not only

aid prediction modeling but also aid exploratory analysis for discovering factors that lead to differential clinical outcomes. Therefore, this dissertation presents novel quantization based WSI representation methods that capture biological variability in patient's samples.

Informative Feature Selection and Reduction

Identification of informative image features is necessary for WSI prediction modeling. Dimensionality reduction in WSI prediction modeling is beneficial for the following reasons: 1) prediction modeling after dimensionality reduction can result in simpler models with higher prediction performance and 2) dimensionality reduction can provide insights about the data by highlighting important features or dimensions [90].To identify informative and robust image features, one of two techniques is generally applied: (1) feature selection or (2) feature reduction. These techniques reduce the dimensionality of the feature space by removing irrelevant and redundant features to improve the performance of prediction modeling.

Feature selection methods can be broadly classified into three categories: filter, wrapper and embedded methods [91]. Filter methods include univariate methods that filter features based on statistical properties (e.g., t-test, Wilcoxon rank sum test, ANOVA, and chi-square) [92]as well as multivariate methods that consider the effects of multiple interacting features (e.g., minimum redundancy maximum relevance (mRMR) [93], and relief-F [94]). Because filter methods are fast and scalable to high-dimensional data, they are often used in pathology informatics [4, 52, 54, 58]. However, filter methods select features independent of the classifier; as such, they may not select optimal feature

sets for a particular classifier. In contrast, wrapper methods generate various subsets of features using a deterministic or randomized search method and directly evaluate them with a classifier. Common wrapper methods are often coupled with a search method and include sequential forward search (SFS) [95], sequential backward elimination (SBE),randomized hill climbing [96], genetic algorithm [97], and simulated annealing [98]. Sequential search methods are commonly used in pathology informatics systems [24, 37, 99]. The drawbacks of wrapper methods include over-fitting and computational cost. Thus, embedded methods identify important features as intrinsic properties of a classifier (e.g., the weight vector of a SVM classifier [100] and the nodes of a random forest or tree classifier [101]). DiFranco et al. used random forest feature selection in their system for detecting regions of prostate tumor in WSIs [73].

Feature reduction techniques transform high-dimensional data into meaningful low-dimensional data. Ideally, reduced dimensionality should correspond to intrinsic dimensionality of data. In comparison to feature selection methods, feature reduction methods transform the original features instead of selecting an optimal feature subset. Moreover, they are unsupervised with the exception of linear discriminant analysis (LDA) methods. Feature reduction methods can be divided into two groups: (1) Linear feature reduction techniques (e.g., principal component analysis (PCA), independent component analysis (ICA), factor analysis, and LDA) and (2) nonlinear feature reduction techniques including multidimensional scaling (MDS), ISOMAP, kernel PCA, local linear embedding (LLE), Laplacian Eigenmaps, and graph embedding[66, 102]. Because of the intuitive interpretation of PCA transformed features, it is one of the most

commonly used feature reduction techniques in pathology informatics [32, 39, 99]. Besides PCA, researchers have also used graph embedding [55], ISOMAP [85], and MDS [103] for feature transformation in pathology informatics systems.

Decision Making

Decision making is the final and an important step for pathology imaging informatics, which develops prediction models for a number of prognostic clinical variables such as cancer grade, cancer subtype, survival time, disease recurrence, and therapeutic response. Prediction models are developed by training a classifier using the knowledge from WSIs and clinical labels. Classification methods commonly used in pathology informatics include k-nearest neighbors (k-NN) [24, 30, 37-39, 71, 85], support vector machines (SVM) [24, 30, 33, 35, 37, 39, 50, 51, 54, 55, 58, 59, 61, 85, 99], Bayesian methods [24, 25, 30, 36, 37, 39, 50, 52, 57, 85, 99], neural networks [63, 104], decision trees[31, 85] and logistic regression[31]. Researchers often evaluate image features using multiple classifiers and report the best-performing classifiers [24, 30, 37, 39, 85]. In addition to basic classifiers, researchers in pathology informatics use boosting algorithms (i.e., combining a weighted set of weak classifiers to produce a robust classifier [25, 71, 85]) and ensemble methods that combine the decisions of multiple classifiers [24]. It is important to note that feature selection/reduction and classification should be conducted within a cross-validation (CV) framework, especially when evaluating systems for clinical prediction [105].

Since the predictive accuracy of supervised learning methods depends on the quality of the training data, researchers are investigating methods for collecting and combining training data for predictive modeling. Training data can be collected using the following methods: (1) one-time annotation by a single pathologist, (2) one-time annotation by multiple pathologists, and (3) run-time continuous annotation. Most informatics systems use the first method. However, the performance of these systems is subjective to the pathologist's knowledge. The second method for annotation requires a method for combining annotations from multiple experts [9]. The third method of annotation falls in the field of active learning or relevant feedback, where one or more pathologists provide active feedback to the learning algorithm in order to iteratively improve its knowledge [9, 106-108]. Although active learning based algorithms may need a longer training phase, they have the potential to evolve into useful CDSSs for clinical applications.

Commercial Systems

The importance of quantitative and objective analysis of biopsy WSIs has led to several commercial software tools for WSI analysis. GENIE™ (Aperio, Vista, CA, USA) separates and quantifies different biological regions and tissue structures in WSIs based on examples provided by pathologists. HALO™ (Indica Labs, Corrales, NM, USA) features fast processing of WSIs, segmentation of tissue structures, and quantification of various properties. AQUAAnalysis™ (HistoRx, Branford, CT, USA) localizes and quantifies protein biomarkers in cellular and sub-cellular regions. Visiopharm (Hoersholm, Denmark) provides three image-analysis tools targeted

specifically for WSI analysis: VisiomorphDP™, TissuemorphDP™, and HER2-CONNECT™. VisiomorphDP™ and TissuemorphDP™ (1) quantify various image properties, (2) use these image properties and user-selected examples of tissue structures to develop a decision rule for classifying/segmenting tissue structures, and (3) measure and report properties of these tissue structures in a large batch of images. HER2-CONNECT™ is a diagnostic tool for scoring HER2-stained, breast-cancer biopsy sections. Aperio has designed an open-architecture solution, called PRECISION, which integrates various commercial tools. All of these tools provide limited image processing capabilities for a complete WSI or for a ROI in a WSI. In most cases, pathologists manually select the ROIs and make diagnoses based on feedback from these commercial tools. Usually, an expert user calibrates these systems for each laboratory-specific experimental setup. To the best of our knowledge, none of these tools provides complete data analysis for clinical decision-making that includes the following steps: (1) automatic quality control of image artifacts and ROI selection in WSIs, (2) automatic adjustment for color and scale batch effects in images collected at multiple institutions, (3) comprehensive information extraction using different types of pixel-, object- and semantic-level image features, (4) knowledge modeling, and (4) predictive models for patient-level diagnosis such as grade, subtype, and prognosis.

Dissertation Structure

This dissertation is motivated by the need to find solutions for two limitations of existing CDSSs for histopathological images: (1) Human intervention is required for region-of-interest (ROI) selection and CDSSs operation and (2) The performance of CDSSs is not robust and is sensitive to data variance [1, 2]. The development of robust CDSSs for WSIs (without prior ROI selection) faces several informatics challenges: (1) Lack of robust segmentation methods for histopathological images, (2) Semantic gap between quantitative information and pathologist's knowledge, (3) Lack of batch-invariant imaging informatics methods, (4) Lack of knowledge models for capturing informative patterns in large WSIs, and (5) Lack of guidelines for optimizing and validating diagnostic models. With the goal of addressing these challenges, this dissertation has three specific aims:

Specific Aim 1: To extract information from histopathological images using robust, batch-invariant, and comprehensive image feature extraction methods.

Specific Aim 2: To model high-level knowledge in WSIs for biological interpretation and decision making.

Specific Aim 3: To validate the power of imaging informatics algorithms for assisting in the diagnosis of histopathology-based endpoints and the prediction of clinical endpoints.

Figure 5 illustrates the structure of this dissertation linking the chapters to the three specific aims. The left to right flow diagram on the top illustrates the flow of information across three specific aims. The top to bottom flow diagrams illustrate specific areas investigated to achieve each specific aim. Quality control algorithms insures that CDSS only processes the artifact-free regions of WSIs. Information extraction algorithms extract quantitative information (image features) from the quality-controlled regions. Knowledge modeling algorithms uses the information to capture morphological patterns

in WSIs as quantized representations. Finally, decision making algorithms uses the knowledge models to develop prediction models for clinical diagnosis and prognosis.

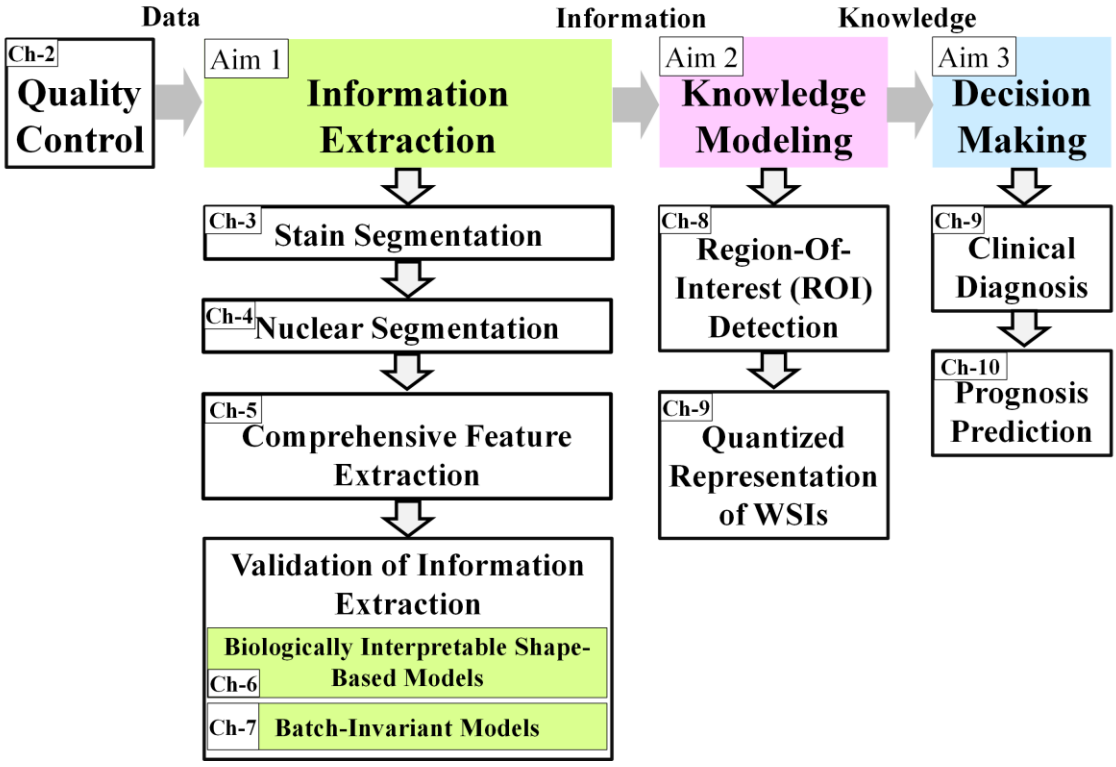

Figure 5: Structure of Dissertation.
Left to right flow illustrates information flow along three specific aims, where quality-controlled regions in WSIs are converted into quantitative information (image features), then the information is used to model biological knowledge in WSIs, and then the knowledge models are used for decision making. Top to bottom flow summarizes specific topics covered for each specific aim.

Chapter 2, 3, and 4 address the informatics challenge of developing robust image segmentation methods. Chapter 2 focuses on tissue-fold artifact detection in WSIs and develops an adaptive tissue-fold segmentation method, ConnSoftT. ConnSoftT calculates optimal segmentation thresholds by modeling the connectivity of tissue structures at various thresholds. Chapter 2 provides quantitative and visual comparison of ConnSoftT to existing methods and highlights pitfalls of existing methods. The adaptive estimation of optimal thresholds using image connectivity rather than intensity can be useful for other segmentation applications. Chapter 3 focuses on segmentation of the color enhanced cellular structures in a stained tissue image and develops a supervised segmentation method. The method (1) incorporates knowledge from pre-segmented reference images for training segmentation models and (2) normalizes new test images to the reference images for batch-invariant performance. Results on four batches of H&E-stained histopathological images indicate that the performance of the supervised method is comparable to user-interactive segmentation. This supervised segmentation system can be easily trained and applied for the segmentation of microscopy images stained with other staining protocols. Chapter 4 focuses on segmentation of dense nuclear clusters in a binary nuclear-stain mask of a tissue image. The proposed method is an edge-based method including the following three steps: concavity detection, straight-line segmentation, and ellipse fitting. The chapter illustrates quantitative performance of the method on simulated data (randomly generated overlapping elliptical shapes) and visual examples on H&E-stained histopathological images. High performance on simulated data indicates that this method will be useful in segmenting elliptical shapes from complex structures of overlapping ellipses.

Chapter 5 and 6 address the informatics challenge of reducing semantic gap by developing interpretable image features. Chapter 5 focuses on the development of a comprehensive set of image features. This comprehensive set includes 12 high-level

feature subsets that represent different aspects of image such as nuclear texture, cytoplasmic shape, and color. The chapter illustrates the utility of the set in decision making for a variety of binary renal tumor subtyping and grading endpoints. For each endpoint, the chapter highlights emergent feature subsets, which can be interpreted biologically. This feature set (with high-level subsets) is a powerful tool for interpreting diagnostics models and discovering imaging markers for variety of cancer endpoints. Chapter 6 focuses on the development of biologically interpretable shape-based features. The proposed shape-based features quantify the distribution of shape patterns in an image using Fourier shape descriptors. Using a case-study on renal tumor subtyping, the chapter compares the prediction performance of novel shape-based features to traditional image features and discusses biological interpretability of differentially expressed shapes in binary subtyping models. Proposed shape-features can be used for describing images with multiple shapes rather than describing individual shapes.

Chapter 7 addresses the informatics challenge of developing batch-invariant informatics methods. It focuses on information-level batch-effects that affect the prediction performance of CDSSs based on images from multiple institutions or set-ups. The chapter develop two categories of batch-effect removal methods: (1) image-level, scales images using a nuclear-area model and (2) information-level, normalizes the distribution of features across batches using parametric or non-parametric feature models. Using four renal tumor histopathological datasets, acquired during different experimental setups, the chapter illustrates the impact of batch effects on image features and downstream cancer predictions. Results indicate that information-level normalization methods can drastically improve cross-batch performance.

Chapter 8 addresses the informatics challenge of finding region-of-interest in large WSIs. The chapter illustrates a visualization tool, called TissueViz, which facilitates the

study of spatial patterns in WSIs using a three-mode design framework: single feature variation, unsupervised multi-dimensional clustering, and supervised classification. Three case-studies on OvCa illustrate the usefulness of TissueViz in facilitating the discovery and biological interpretation of quantitative image features and the identification of ROIs in WSIs.

Chapter 9 addresses the informatics challenge of developing high-level representation for WSIs that can model pathologists' knowledge. High-level representations are essential to tackle biological variation in WSIs while making diagnostic decisions. Instead of finding optimal ROIs for each cancer endpoint, the chapter proposes representing different biological regions in WSIs using quantized representations including univariate, multivariate, and multivariate subset quantization. Case studies on binary KiCa and OvCa endpoints illustrate the effect of different WSI-representation strategies and ROI selection on prediction performance. The chapter validates these representation strategies for two applications: (1) diagnosis of histopathology-based endpoints such as subtype and grade and (2) prediction of clinical endpoints such as metastasis, stage, lymphnode spread, and survival. The quantized representation of WSIs allows data mining methods to extract informative biological regions while training models for different cancer endpoints.

Chapter 10 addresses the informatics challenge of optimizing and validating prognosis prediction models. It discusses the development of an interactive patient-level visualization tool, PatientViz, which allows user to study patient stratification in terms of prognostic significance, stability, and reproducibility simultaneously. The chapter also develops a method for genomic stratification using histopathological knowledge. A case-study on prognostically significant KiCa patient stratification illustrates the usefulness of histopathological knowledge as compared to clinical five-year-survival labels.

Chapter 11 concludes the dissertation with a discussion of contribution to the field of pathology imaging informatics. Also, the chapter provides an outlook about the future directions of research and open challenges.

CHAPTER 2

ADAPTIVE SEGMENTATION OF TISSUE-FOLD ARTIFACTS IN WHOLE-SLIDE IMAGES

Introduction

An important challenge in pathology imaging informatics is developing robust image segmentation methods that can overcome various biological and technical diversities. An image segmentation method may be designed to segment different objects in WSIs such as tissue regions, artifacts, stains, and nuclei. This chapter focuses on tissue-fold artifact segmentation while Chapter 3 and Chapter 4 will focus on histopathological stain and nuclear segmentation, respectively. The research presented in this chapter was conducted in collaboration with other researchers and most of the content is part of a published article on tissue-fold artifact detection [109].

The presence of artifacts in histopathological images affect the downstream image features and decision models in CDSSs [11, 16]. Image artifacts such as tissue folds, out-of-focus regions, and chromatic aberrations are often found in digital images of tissue-biopsy slides. Among these artifacts, the occurrence of out-of-focus regions and chromatic aberrations can be prevented during the image acquisition stage using advanced microscopes. In contrast, the occurrence of tissue folds cannot be easily prevented during slide preparation, when a thin tissue slice folds on itself. Therefore, while studying a biopsy slide under a microscope, pathologists avoid tissue regions with folds. Similarly, CDSSs must also detect and avoid tissue-fold regions.

In recent pathology imaging informatics studies involving WSIs, researchers avoid tissue folds by manually selecting images or ROIs [65, 110]. Even though manual selection ensures the quality of selected tissue regions, it limits the speed and objectivity of computer-aided analysis by introducing a user-interactive step and by adding user subjectivity. Moreover, manual selection is a tedious process for large datasets. For example, datasets from TCGA include cancer endpoints with more than a thousand WSIs, most of which have tissue folds [3].

Recent studies have proposed methods for detecting tissue folds. Palokangas *et al.* proposed an unsupervised method for tissue-fold detection using k-means clustering [17]. This method detects most of the prominent folds if a variety of folds are present on the slide. However, the method fails if no folds are present (i.e., the method assumes that folds are present, resulting in false positives). To detect tissue folds, Bautista and Yagi proposed a color-based method with a fixed threshold [18, 111]. Unlike unsupervised clustering, this method does not fail for WSIs without folds. However, a fixed threshold is not effective for all WSIs, especially if there are data batch effects (e.g., images acquired with different microscopes). Although the two methods differ with regard to how they detect tissue folds, the studies agree about the utility of color saturation and intensity properties for tissue-fold detection in low-resolution WSIs.

This chapter proposes a novel method for detecting tissue folds and compare the method to other tissue-fold detection methods. The proposed method detects tissue folds in low-resolution WSIs using an adaptive soft-threshold technique in which two

thresholds—soft and hard—are determined using a model based on the connectivity of tissue structures at various thresholds. The threshold model is trained on a set of manually annotated tissue folds. The two thresholds are then used in conjunction with a neighborhood criterion to find tissue folds. We test the proposed method on a separate set of manually annotated test images. We also compare our method to two other methods: an unsupervised clustering-based method proposed by Palokangas *et al.* and a simplified form of our supervised method, which optimizes two thresholds directly from the train set instead of using a connectivity-based model.

Materials and Methods

Datasets

We use publicly available WSIs of H&E-stained tumor samples of OvCa and KiCa provided by TCGA [3]. TCGA provides the WSIs of tumor samples at four different resolutions. We use the lowest-resolution image for tissue-fold detection because images at the lowest resolution are much easier to load and faster to process. Moreover, tissue folds are distinctly visible at the lowest resolution.

We evaluate the performance of tissue-fold detection using a set of 105 manually annotated images for each cancer endpoint. We annotate all tissue-fold regions in a WSI by clicking points on the boundary of every fold and enclosing it within a polygon. Figure 6 illustrates examples of WSIs for tumor samples from OvCa and KiCa patients with manually annotated tissue folds.

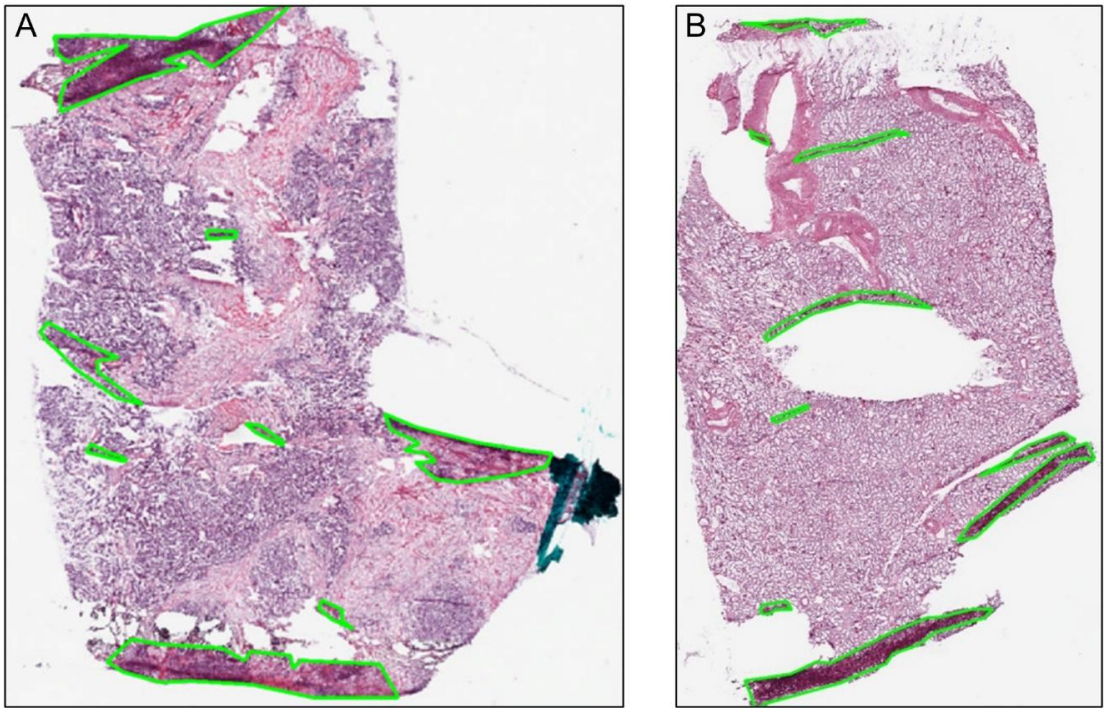

Figure 6: Manual annotation of tissue folds in WSIs.
Tissue folds marked in WSIs of two types of carcinomas: (A) ovarian serous adenocarcinoma (OvCa) and (B) kidney renal clear cell carcinoma (KiCa).

Tissue-Region Identification

Before detecting tissue folds in a WSI, we identify the regions of tissue. A typical TCGA WSI contains large white regions representing blank, tissue-less portions of the slide and some bluish-green regions representing pen marks used by pathologists to annotate the slide. These blank and pen-marked regions are not informative for cancer diagnosis, so we remove these regions from further consideration. For convenience, we represent the tissue, blank, and pen-marked regions as logical matrices A, W, and P, respectively, with dimensions equal to the WSI dimensions. The value of **A**, **W**, and **P** at a pixel location (x,y) is given by $a(x,y)$, $w(x,y)$, and $p(x,y)$, respectively. We use hue (h), saturation (s), and intensity (i) of the pixel (x,y) to determine w and p, given by

$$w(x,y) = \mathrm{II}[s(x,y) \le 0.1] \tag{1}$$

$$p(x,y) = \mathrm{II}[(0.4 < h(x,y) < 0.7 \wedge s(x,y) > 0.1) \wedge i(x,y) < 0.1], \tag{2}$$

where II(c) is an identity function that returns logical 1 if c is true. We classify pixels with no color (i.e., saturation less than 0.1) as part of a blank region; and we classify pixels with either bluish-green (i.e., a hue between 0.4 and 0.7 and saturation greater than 0.1) or black (i.e., intensity less than 0.1) as part of the pen-marked region. We empirically found that pen-marks in TCGA WSIs are either bluish-green or black. In addition to pen marks, noise occurs in the **P** mask. Thus, we remove all connected regions with an area of less than five pixels. Similarly, the **W** mask, in addition to blank regions, contains white, no-stain-tissue regions such as glands of tissue. We morphologically open the W mask to isolate these regions from the blank region and then remove them using an area threshold. Because the scaling factor between the thumbnail and the largest resolution varies in TCGA with the size of the WSI, we use an adaptive

area threshold equivalent to the area of a tile (512x512 pixels) in the highest resolution of the WSI. Finally, if a pixel is zero in both the W and P masks, then it is one in tissue mask **A,** given by

$$a(x,y) = \mathrm{II}\{\neg[w(x,y) \cup p(x,y)]\}. \quad (3)$$

We use only tissue regions for fold detection. Figure 7 is an example result for the detection of tissue regions. In Figure 7.B, we have painted the pen-marked and blank regions as gray and black, respectively.

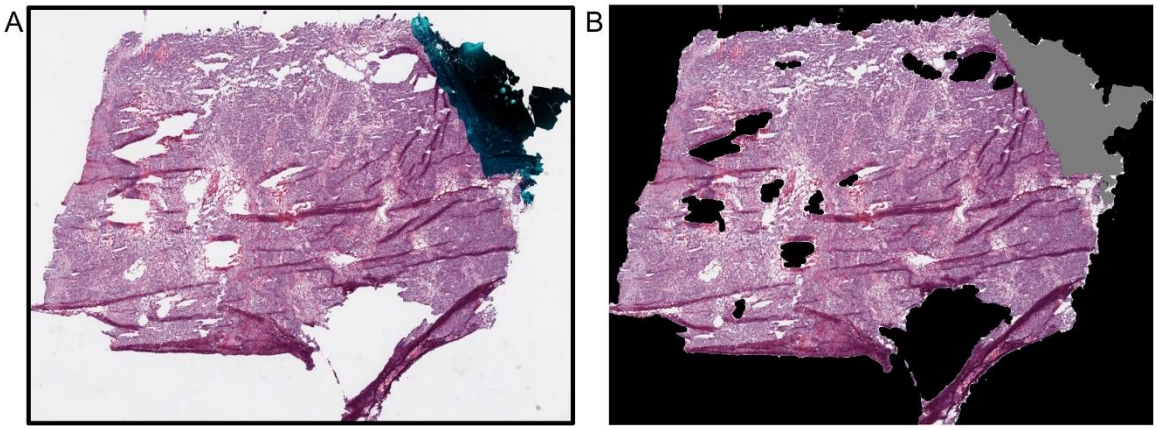

Figure 7: Tissue-region detection in a WSI.
(A) Original RGB thumbnail and (B) painted thumbnail, in which pen-mark and blank regions are painted gray and black, respectively

Tissue-Fold Detection

Connectivity-based soft threshold (ConnSoftT)

We propose a novel method for detecting tissue folds in WSIs by exploring the color and connectivity properties of tissue structures. A WSI of a tissue-biopsy slide stained with H&E has three primary regions: tissue structures (nuclei and cytoplasm), blank slide, and tissue folds. These regions differ in their color saturation and intensity properties: (1) Tissue folds are regions with multiple layers of stained tissue resulting in image regions with high saturation and low intensity [17], (2) nuclear regions are stained blue-purple and have low intensity, and (3) cytoplasmic regions are stained pink and have high intensity. Therefore, we apply color saturation and intensity values to classify a pixel (x, y) into the tissue-fold region. We subtract the color intensity $i(x, y)$ from color saturation $s(x, y)$ for each pixel, resulting in a difference value $d(x, y) = s(x, y) - i(x, y)$, where $d(x, y) \in [1, -1]$. Typically, $d(x, y)$ is high in tissue-fold regions, intermediate in nuclear regions, and low in cytoplasmic regions. If we threshold the difference image, **D** (including all pixels), with various thresholds in the range of negative one to one, $t \in \{-1, -0.95, .., 0, .., 0.95, 1\}$, then the following three patterns emerge: (1) At high thresholds, only a few connected objects (i.e., mostly tissue folds) are segmented; (2) at medium thresholds, a large number of connected objects (i.e., mostly tissue folds and nuclei) are segmented; and (3) at low thresholds, only a few large connected objects (i.e., tissue structures merged with cytoplasm) are segmented. Our goal is to find an optimal threshold that segments only tissue folds. However, we observed that this threshold varies because of variations in tissue samples, preparation sites, and acquisition systems. Thus,

we hypothesize that an approximate tissue-fold threshold can be predicted based on object connectivity in a WSI.

In Figure 8, we illustrate the following for a WSI: (1) the difference image, (2) manually annotated folds, (3) segmented binary images $\mathbf{B}(t)$ at various thresholds, and (4) the distribution of connected-object count $C(t)$ using 8-connectivity at various thresholds. The peak of the distribution corresponds to approximate threshold at which dark nuclear structures are segmented but not merged. We hypothesize that tissue folds can be detected by a threshold greater than the threshold corresponding to the peak, and this threshold is a function of connected-object count at the peak. Our hypothesis is based on an assumption that, for any dataset, tissue-fold objects are a small percentage of all connected objects at the peak. We can safely make this assumption because of the nature of tissue folds. Tissue-fold artifacts are caused by the folding of tissue slices when placed on a glass slide. Thus, folds are seldom randomly distributed over the whole-slide, which would lead to a large number of connected objects (greater than the number of nuclear objects). Even if a large portion of the image contains tissue folds, most tissue-fold pixels are likely to be connected within a small number of tissue-fold regions.

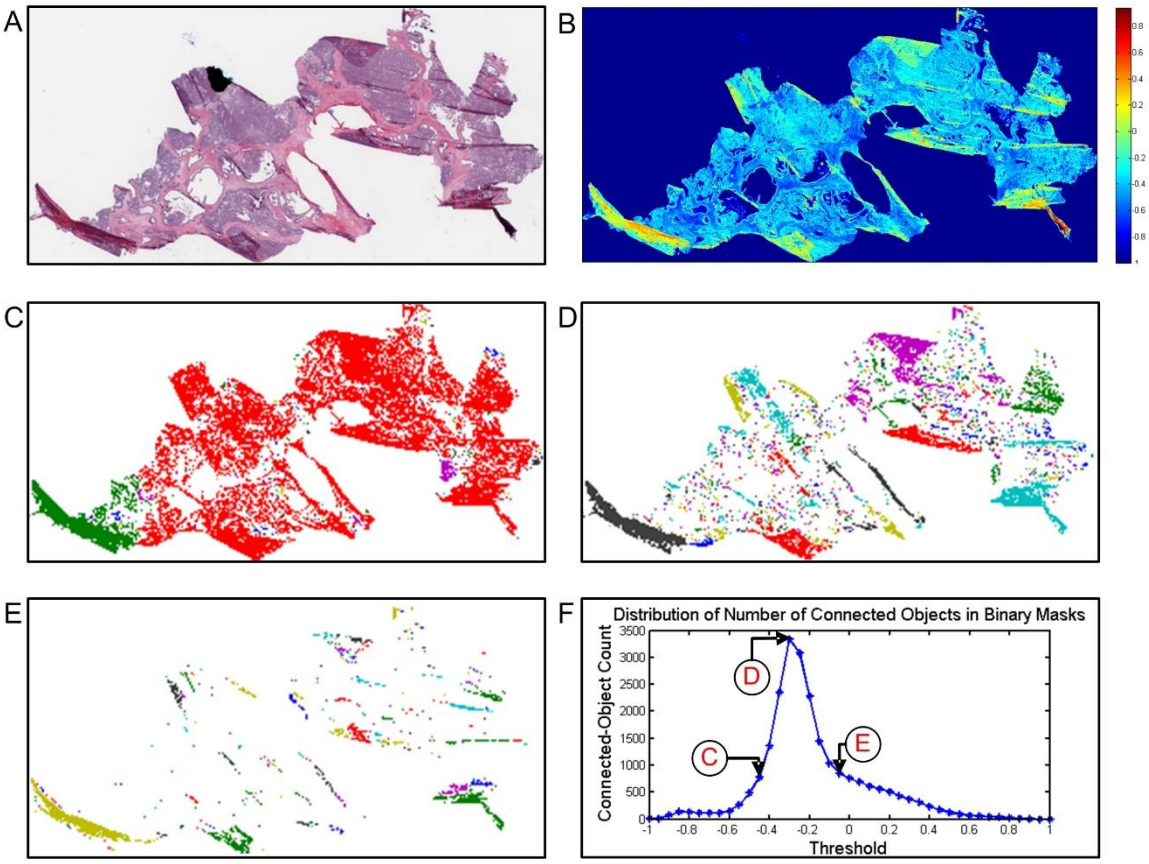

Figure 8: Estimation of soft and hard thresholds for detecting tissue folds using the connectivity-based soft threshold (ConnSoftT) method.
We illustrate threshold estimation using an example OvCa WSI (A), which has multiple tissue folds, detected by manual annotation, and indicated in the binary mask (B). Using the saturation and intensity of an OvCa WSI (A), we calculate a difference image (saturation-intensity) (C) and then threshold the difference image to generate binary masks at various thresholds, including -0.45 (D), -0.3 (E), and -0.05 (F). The connected objects in the binary masks are randomly pseudo-colored to highlight separate objects. We count the number of connected objects in all binary masks to estimate a distribution (G); and then we use this distribution to calculate the optimal thresholds. For parameters α=0.64 and β=0.34, the optimal thresholds for this image are t_{hard} = -0.15, and t_{soft} = -0.2.

The difference value, $d(x, y)$, for tissue folds in a WSI varies within a range, especially in the area surrounding a strong tissue fold. Therefore, we propose using two thresholds—hard, t_{hard}, and soft, t_{soft}—and a neighborhood criterion. Both thresholds are a function of connected-object count $C(t)$ given by

$$t_{hard} = \mathrm{T}\left(\alpha * \max_{t} C(t)\right) \tag{4}$$

$$t_{soft} = \mathrm{T}\left(\beta * \max_{t} C(t)\right), \tag{5}$$

where T is a function of count defined by $T(c) = \max\{t \mid C(t) \geq c\}$. Based on these thresholds, we classify a pixel as a tissue fold if the following conditions are true: (1) It has a difference value, $d(x, y)$, higher than the soft threshold and (2) it is in the 5x5 neighborhood of a pixel with a difference value greater than the hard threshold. Mathematically, this is given by

$$f(x, y) = \mathrm{II}\left\{\left[b(t_{soft}, x, y) = 1\right] \wedge \left[\sum_{k=-5}^{5}\sum_{l=-5}^{5} b(t_{hard}, x+k, y+l) > 0\right]\right\} \tag{6}$$

where

$$b(t, x, y) = \mathrm{II}(d(x, y) > t). \tag{7}$$

The soft threshold allows tissue-fold pixels in the neighborhood of strong tissue folds to have a lower difference value. The value of $f(x, y)$ for all pixels generates a tissue-fold image, **F**. However, **F** may still have some small, noisy connected objects. We discard these noisy objects in the tissue-fold image using an adaptive area threshold equivalent to five percent of the area of a high-resolution tile in a WSI (i.e., 512x512 pixels).

The hard and soft thresholds depend on parameters α and β. Both tissue morphology and connectivity differ from one cancer endpoint to another cancer endpoint. Thus, we optimize α and β for each TCGA dataset. We optimize these parameters on a set of training images and then evaluate the selected parameters on a set of testing images. We split our annotated data (105 images for each cancer) into 50 pairs of training and testing sets using ten iterations of 5-fold CV. For each CV split, we select optimal parameters by maximizing the average adjusted Rand index (ARI) on the training set. The Rand index is a statistical measure that quantifies the similarity between two sets of data clusters. When applied to tissue-fold detection, the Rand index counts the number of agreements in pixel pairs between the detected tissue-fold pixels and the ground truth. For example, if both pixels in a pixel pair are part of the same class in the ground truth (i.e., either both pixels are tissue folds, or they are not), a pixel pair agrees with the ground truth if both pixels are detected as being in the same class (e.g., tissue folds). Alternatively, if both pixels in a pair are in different classes in the ground truth (i.e., one pixel is a tissue fold, and the other is not), the pair is not in agreement if both pixels are detected as being in the same class. The Rand index is the ratio of the number of agreeing pairs to the total number of pairs. To account for different class prevalence (i.e., different numbers of pixels in tissue-fold vs. non-tissue-fold regions), the adjusted Rand index is a modification of the Rand index. For two classes, the adjusted Rand index is given by

$$\mathrm{ARI} = \frac{\sum_{g,p}\binom{m_{g,p}}{2} - \sum_{g}\binom{m_{g,\circ}}{2}\sum_{p}\binom{m_{\circ,p}}{2}\Big/\binom{M}{2}^{2}}{\frac{1}{2}\left[\sum_{g}\binom{m_{g,\circ}}{2} + \sum_{p}\binom{m_{\circ,p}}{2}\right] - \sum_{g}\binom{m_{g,\circ}}{2}\sum_{p}\binom{m_{\circ,p}}{2}\Big/\binom{M}{2}}, \tag{8}$$

where $m_{g,p}$ indicates the number of pixels designated as g in the ground truth and predicted to be p, where $(g, p) \in \{fold, tissue\}$f. For example, $m_{fold,tissue}$ indicates the number of pixels designated as part of the tissue-fold regions in the ground truth, but it is predicted to be part of non-tissue-fold regions. M is the total number of pixels; $m_{g,\circ} = m_{g,fold} + m_{g,tissue}$ is the total number of pixels designated as g in the ground truth; and $m_{\circ,p} = m_{fold,p} + m_{tissue,p}$ is the number of pixels predicted to be p. Because a segmented image can be perceived as a clustering of pixels into groups, the Rand index and its various forms are often used for image-segmentation evaluation [112, 113]. We have chosen ARI for our evaluation because it is adjustable based on class prevalence. In most WSIs, tissue-fold regions are a small percent of the tissue region. Thus, errors in fold detection will not significantly affect a metric that is not adjustable for class prevalence. For example, accuracy calculates the number of pixels assigned to the correct class regardless of the class. Since we have more tissue pixels than tissue-fold pixels, it prefers methods that classify tissue pixels correctly even if the methods compromise the performance of fold detection. In other words, metrics that do not account for prevalence tend to severely down-weight the sensitivity of tissue-fold detection.

We optimize α and β in the range of 0 to 1 with two levels of quantization: coarse and fine. While optimizing, we allow only pairs in which α is greater than β so that t_{hard} is greater than t_{soft}. During the coarse optimization, we vary the parameters with steps of 0.1 in the range of 0 to 1 and calculate the parameter pair, α_c and β_c, with the maximum ARI, averaged over all training samples. During fine optimization, we vary the

parameters with steps of 0.01 in the range of α_c-0.1 to α_c+0.1 and β_c-0.1 to β_c+0.1. The two-level optimization speeds up the optimization process.

Clustering (Clust)

As a comparison, we implement a clustering-based method for tissue-fold detection suggested by Palokangas et al. [17]. This method has three steps: preprocessing, segmentation, and the discarding of extra objects. In our implementation, we first detect tissue regions in a WSI and then follow these three steps. First, we subtract smoothed and contrast-enhanced saturation $\hat{\mathbf{S}}$ and intensity $\hat{\mathbf{I}}$ images of a WSI and calculate difference image $\hat{\mathbf{D}}$. Second, we cluster the pixels of the difference image using k-means clustering and assign the cluster of pixels with center at the maximum difference value as tissue folds. Finally, we discard extra objects in the tissue-fold image using an adaptive area threshold equivalent to five percent of a tile area in the highest resolution of the WSI. For k-means clustering, we optimize the number of clusters, n, based on the change in the average sum of the difference (variance) over all clusters. We start optimization with n=2 clusters and terminate at n=6 clusters; we select a value of n for which the change in variance compared to the variance with n-1 clusters is less than one percent.

Soft threshold (SoftT)

Instead of clustering the difference image, we can also find tissue folds by applying a soft and hard threshold, as done in the proposed ConnSoftT method. However, in the ConnSoftT method, we apply adaptive thresholds based on tissue connectivity after optimizing α and β. Alternatively, we can directly optimize the hard, H_T, and soft, S_T,

thresholds for a dataset. Therefore, for comparison, we implement the direct-optimization version for soft thresholding by repeating all the steps from the ConnSoftT method, excluding the connectivity-based analysis and optimization steps. After obtaining the difference image, **D**, we optimize H_T and S_T in the range of -1 to 1, with the condition that the hard threshold is greater than the soft threshold. Similar to the ConnSoftT method, we optimize the thresholds using two quantization levels (i.e., coarse with steps of 0.2 and fine with steps of 0.02), manually annotated training data, and the ARI performance metric. Finally, after thresholding the difference image, we discard noisy objects using an adaptive area threshold (the same threshold as in the ConnSoftT method).

Results and Discussion

Comparison of ConnSoftT, Clust, and SoftT Methods

In this section, we discuss the performance of the ConnSoftT method and compare it to two other methods: Clust and SoftT. We test the methods on two datasets of 105 images with manually-annotated folds of OvCa and KiCa samples. Using ten iterations of 5-fold CV, we divide the datasets into 50 pairs of training and testing sets, in which each training set is used for optimizing models in the ConnSoftT and SoftT methods while the testing set is used to test all three methods. We assess the performance of detecting tissue folds using four metrics: (1) ARI, which was also used for model optimization, (2) the true positive rate (TPR), or sensitivity, (3) the true negative rate (TNR), or specificity, and (4) the average true rate (ATR) (i.e., the average of TPR and TNR). The average and standard deviation of performance metrics over 50 iterations of CV on KiCa and OvCa

images are listed in Table 1 and Table 2, respectively. The best method should result in all metrics closer to one. A high TNR and a low TPR indicates that the method under-segments tissue folds while a low TNR and a high TPR indicate that the method over-segments. Therefore, the best method should have high ATR. From these tables, we can make several observations. First, compared to the other methods, the ConnSoftT method detects tissue folds more effectively because it has the highest ARI (0.50 in KiCa and 0.40 in OvCa) and ATR (0.77 in KiCa and 0.73 in OvCa). Second, based on TNR, TPR and ATR, the Clust method under-segments tissue folds (TNR is highest at 0.99 in KiCa and 0.98 in OvCa) while the SoftT method over-segments tissue folds (TPR is highest at 0.62 in KiCa and 0.57 in OvCa). The ConnSoftT method achieves a balance between the two methods (ATR is highest at 0.77 in KiCa and 0.73 in OvCa). Third, all three methods have lower TPR than TNR. TPR is more sensitive to faults in tissue-fold detection than TNR because the positive class (tissue-fold regions) has a lower prevalence than the negative class (non-tissue-fold regions). The difference in the prevalence is the main motivation for using ARI, which is adjusted for prevalence, for parameter optimization in the ConnSoftT and SoftT methods.

Table 1: Tissue-fold detection performance in KiCa WSIs.

Metric	Clust	SoftT	ConnSoftT
ARI	0.43±0.03	0.39±0.05	0.50±0.03
ATR	0.72±0.01	0.75±0.02	0.77±0.01
TPR	0.45±0.04	0.62±0.05	0.55±0.04
TNR	0.99±0.00	0.88±0.04	0.98±0.00

Table 2: Tissue-fold detection performance in OvCa WSIs.

Metric	Clust	SoftT	ConnSoftT
ARI	0.35±0.03	0.31±0.04	0.40±0.03
ATR	0.70±0.01	0.73±0.02	0.73±0.02
TPR	0.41±0.03	0.57±0.06	0.47±0.04
TNR	0.98±0.01	0.88±0.03	0.98±0.00

ARI: Adjusted Rand index
ATR: Average true rate = (TPR+TNR)/2
TPR: True positive rate, or sensitivity = TP/P
TNR: True negative rate, or specificity = TN/N

Figure 9 illustrates tissue-fold detection results for three WSIs using the three methods with the final model parameters. Since Clust is an unsupervised method, when it finds a cluster of pixels in a WSI with the highest difference value, it is not certain if this cluster represents tissue folds or if this cluster includes all of the tissue-fold pixels in the WSI. Figure 9 presents the results of this uncertainty. Figure 9.C and Figure 9.H show that the Clust method under segments, and Figure 9.M shows that the Clust method over segments. Although the SoftT method is supervised, because of the variations in the color properties of WSIs, fixed thresholds cannot successfully segment tissue folds in all WSIs. For example, Figure 9.A and Figure 9.K depict over-segmentation using the SoftT method when WSIs are darker than the remaining set of images. In contrast, ConnSoftT is supervised, and it adapts to a WSI based on its tissue connectivity, which results in more effective tissue-fold detection regardless of variations across images within a dataset.

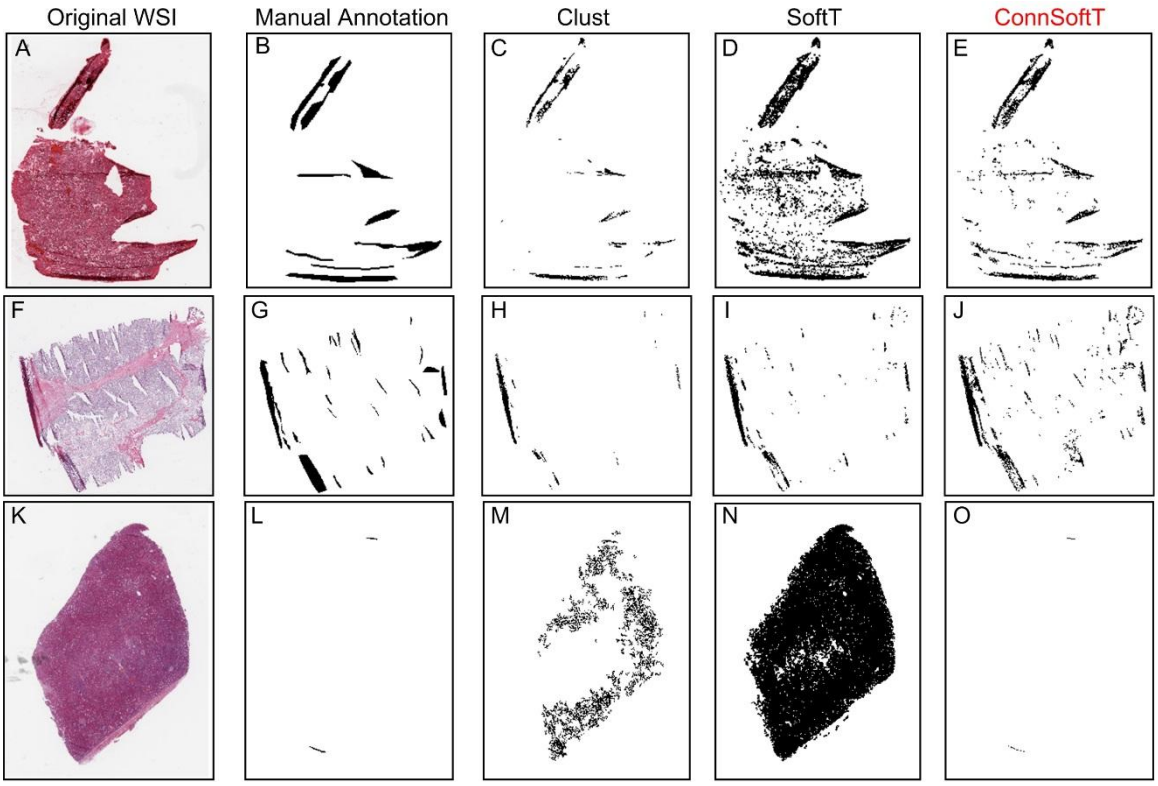

Figure 9: Comparison of the performance of the three tissue-fold detection methods. Tissue folds detected by the three methods: clustering (Clust) (C, H, and M), soft threshold (SoftT) (D, I, and N), and connectivity-based soft threshold (ConnSoftT) (E, J, and O) for an OvCa WSI (A) and two KiCa WSIs (F and K). If tissue folds in a WSI vary in color (A and F), Clust method under segments. On the other hand, if a WSI has no tissue folds in (K), Clust over segments. Because of the fixed thresholding of the SoftT method, it over segments WSIs (A and K) with darker tissue regions and under segments WSIs (F) with lighter tissue folds.

Parameter Optimization and Sensitivity Analysis

In this section, we discuss variation in parameters depending on training samples, adaptive nature of ConnSoftT method, and sensitivity of ConnSoftT method's performance to parameters.

Both ConnSoftT and SoftT methods have two parameters, which are optimized using ground truth of train samples during 50 iterations of CV. In Figure 10, we illustrate the frequency of parameter-pair selection using color maps. For both methods, the optimal parameters are repetitively selected within a local area of the parameter space, which extends from -1 to 1 for H_T and S_T and from 0 to 1 for α and β.

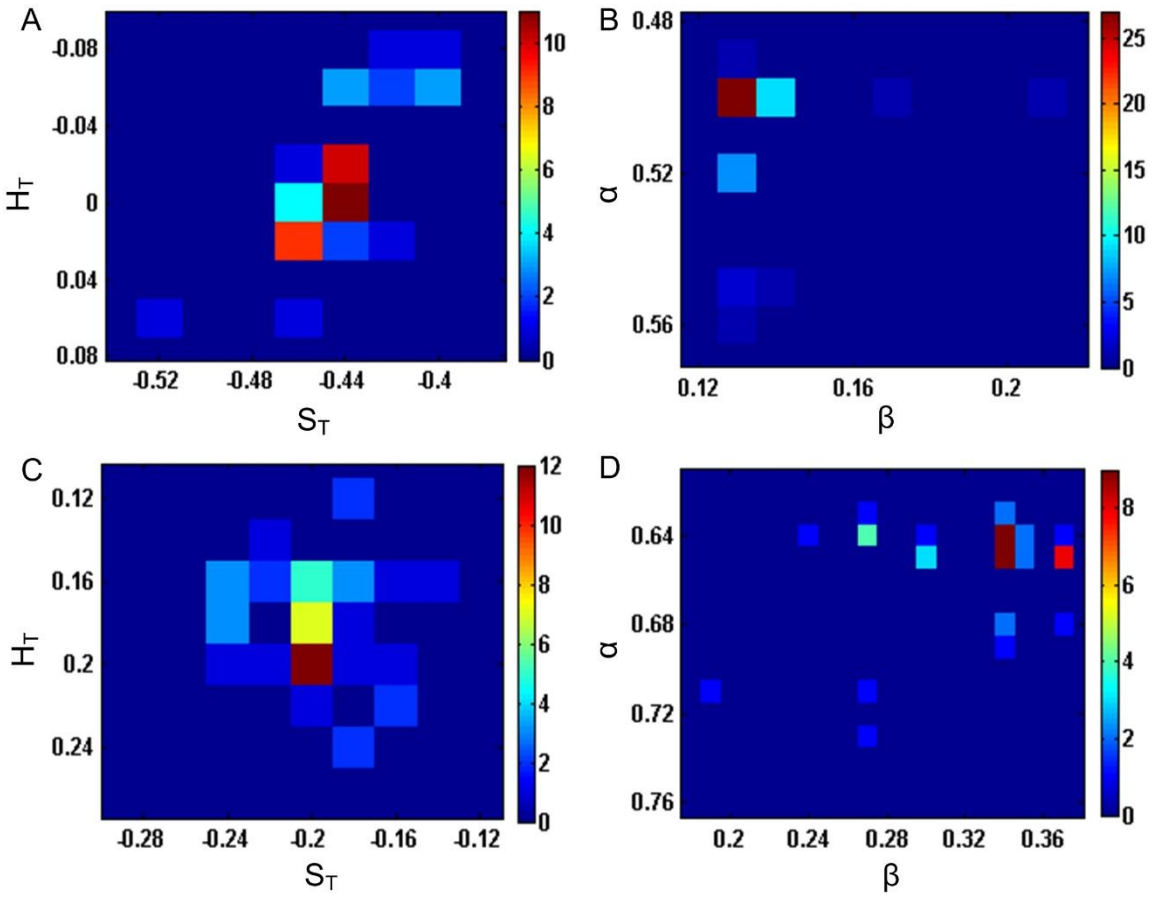

Figure 10: Optimal parameter selection in soft threshold (SoftT) and connectivity-based soft threshold (ConnSoftT) tissue-fold detection methods.
Heat map for the frequency of parameter-pair selection during 50 iterations (5-fold, 10 iterations) of CV for KiCa (A-B) and OvCa (C-D) images. For the SoftT method, the hard and soft thresholds were optimized (A and C). For the ConnSoftT method, α and β were optimized (B and D). Note: In all heatmaps, the parameter space with no selection (zero frequency) has been cropped.

The average of selected parameter pairs during CV (Table 3) closely resembles the parameters of the final models (Table 4), which were optimized using the complete set of 105 images. Low standard deviation in the selection of the parameter pairs during CV and the similarity of average parameters to the final model parameters indicate that the selection of parameters (in both methods) is robust to variation in training samples for a cancer endpoint. Moreover, the difference in the optimal parameters of the two cancer endpoints supports our hypothesis that the pair α and β should vary from one cancer endpoint to another because of differences in morphology between the endpoints.

Table 3: Mean and sample deviation of selected parameters of SoftT and ConnSoftT tissue-fold detection methods during CV.

Data	**SoftT**		**ConnSoftT**	
	$\mathbf{H_T}$	$\mathbf{S_T}$	**α**	**β**
KiCa	-0.01 ± 0.032	-0.44 ± 0.020	0.51 ± 0.015	0.13 ± 0.012
OvCa	0.18 ± 0.026	-0.20 ± 0.023	0.65 ± 0.020	0.33 ± 0.039

Table 4: Parameters in final models of SoftT and ConnSoftT tissue-fold detection methods estimated using the entire datasets.

Data	**SoftT**		**ConnSoftT**	
	$\mathbf{H_T}$	$\mathbf{S_T}$	**α**	**β**
KiCa	0	-0.44	0.5	0.13
OvCa	0.18	-0.20	0.65	0.34

For optimal segmentation of tissue folds, hard and soft thresholds should adapt for every image within a cancer endpoint. In the ConnSoftT method, thresholds depend on parameters (α and β) and the connected-object function. The function adapts for each image to give the optimal segmentation. For instance with the final α and β parameters (Table 4), soft and hard thresholds for 105 OvCa WSIs are within the ranges 0.0995±0.2101 and -0.0162±0.1981, respectively. Similarly, soft and hard thresholds for 105 KiCa WSIs are within the ranges -0.0724±0.2234 and -0.2586±0.2139, respectively. In contrast, the soft and hard thresholds are fixed in the SoftT method (Table 4). Hence, the connectivity-based method successfully adapts for each image compared to the SoftT method.

Figure 11 illustrates the sensitivity of the ConnSoftT method to α and β parameters using performance heatmaps. The heatmap illustrates average ARI for tissue-fold detection on the entire data set of 105 images using the allowable set of coarse parameters ($\alpha > \beta$). The performance of the method is quite similar in the range of parameters selected during CV (marked by dashed rectangle). Thus, tissue-fold detection is not sensitive to small changes in α and β.

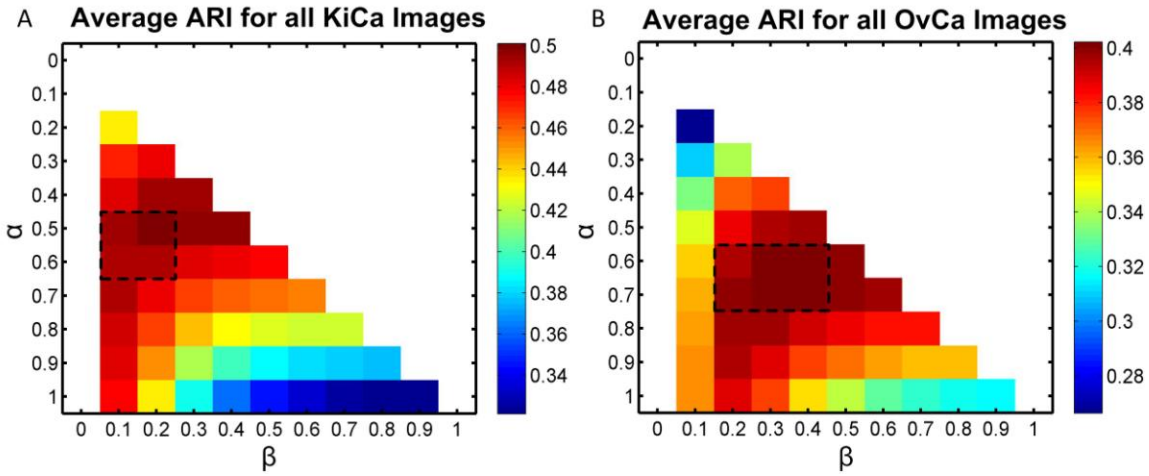

Figure 11: Sensitivity of the performance of connectivity-based soft threshold (ConnSoftT) method to parameter selection.
Heatmap for the average performance (ARI) of tissue-fold detection using Connsoft Method with different parameters. The average was calculated using the entire data set of 105 images for both KiCa (A) and OvCa (B). The performance of the method is quite similar in the range of parameters (marked by a dashed rectangle) selected during CV (Figure 10), indicating that tissue-fold detection is not sensitive to small parameter changes.

ConnSoftT Performance in WSIs with Multimodal Connected Component Distribution

Certain WSIs may result in a multimodal connected-component distribution when thresholding the difference image. In our datasets, a WSI had a multimodal distribution in the following conditions: (1) if difference in stain intensities between tumor and non-tumor regions of WSI (Figure 12.A and Figure 12.E), (2) if non-tissue regions of the WSI are not completely eliminated during the preprocessing step (Figure 12.I and Figure 12.M), and (3) if multiple tissue sections are present on a slide with color variations (Figure 12.Q). Among these conditions, first two are more likely to occur while third condition is very rare. We found only one instance of third condition among all the WSIs considered in the study. In the first condition, a small peak is formed very close the main peak. ConnSoftT method finds all folds in tumor region but miss some folds in non-tumor regions (Figure 12.H). In the second condition, a large peak is formed far left of the main peak and the performance of the method is mostly unaffected. The last condition essentially results in a combined distribution for two images and all tissue-folds in the darker image are detected but some in the lighter image are missed (Figure 12.T).

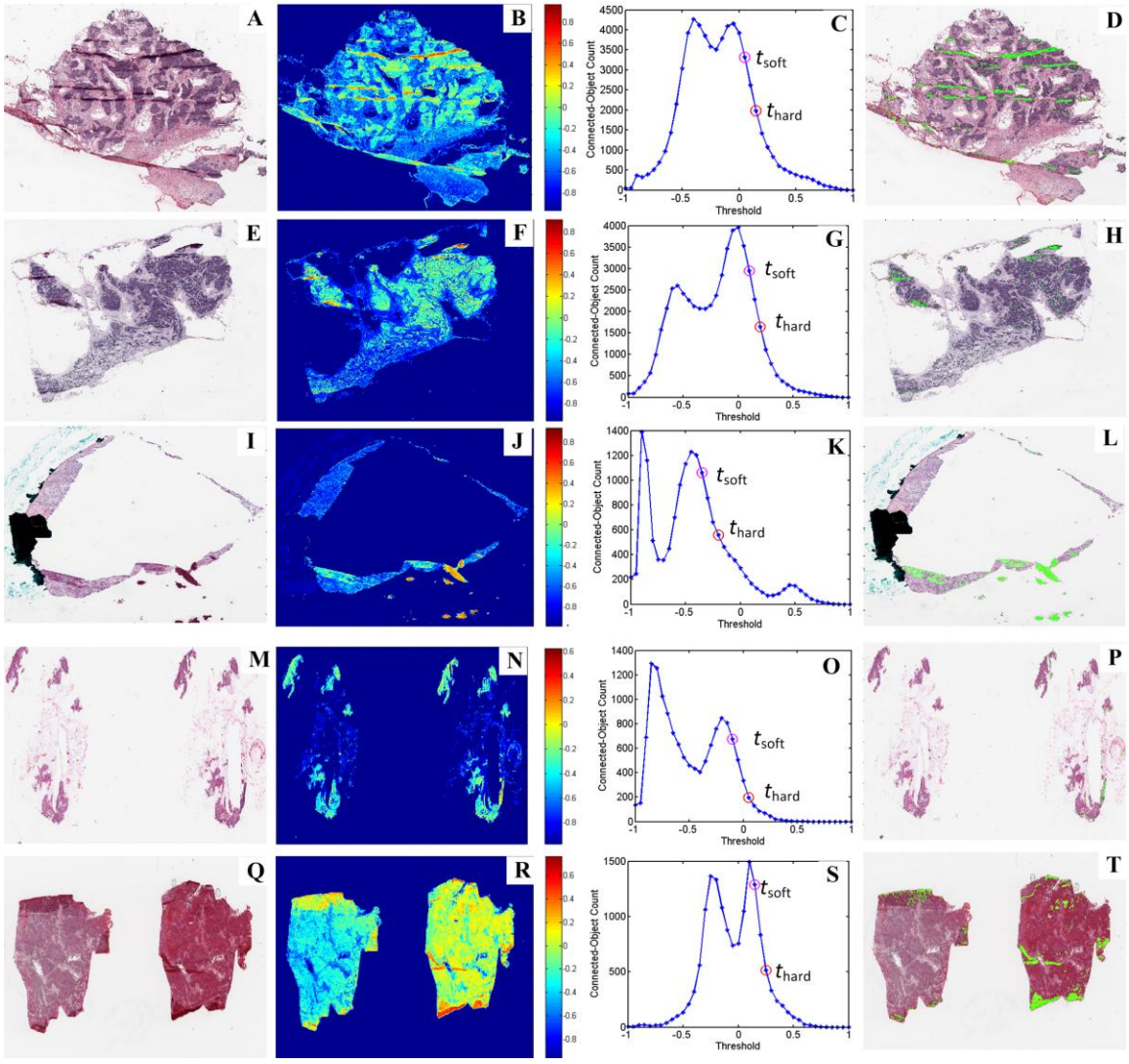

Figure 12: Tissue-fold artifact detection in WSIs with multimodal connected-component count distributions

(A, E) Non-tumor regions have different stain intensities and form a separate peak slightly left of the main peak, (I, M) non-tissue regions (e.g., pen-marks and inadequately stained portions) are not eliminated in the pre-processing step and form a separate peak far left of the main peak, and (Q) Two separate tissue portions with large difference in stain colors form two separate peaks. From left to right, the columns represent original WSI, difference image (S-I), connected-component distribution, and WSI with marked tissue folds. For each WSI, the soft and hard thresholds estimated by the method are marked on the distribution.

Conclusion

Pathology imaging informatics methods strive to develop robust image segmentation methods that can overcome biological and technical challenges posed by histopathological WSIs. In this chapter, we developed a novel image segmentation method, which adaptively estimates soft and hard thresholds based on object connectivity in saturation and intensity color space of an image. We applied this method for detecting tissue-fold artifacts from low-resolution WSIs. Compared to two other methods, our method performed better based on the adjusted Rand index and the average true rate. Tissue-fold artifact detection is essential for insuring only high-quality tissue portions of a WSI are used for downstream diagnosis.

CHAPTER 3

BATCH-INVARIANT SUPERVISED SEGMENTATION OF HISTOPATHOLGICAL STAINS

Introduction

Similar to Chapter 2, this chapter also focuses on developing a robust image segmentation method for histopathological images. The objective of the segmentation method developed in this chapter is to segment stains in histopathological images. The research presented in this chapter was conducted in collaboration with other researchers and most of the content is part of a published article [114].

Color-enhanced, or stained, cellular structures in histological images enable clinicians to identify morphological markers of a disease, and to proceed with therapy accordingly. However, because of variations in specimen preparation, staining, and imaging, resulting images may exhibit very different colors (Figure 13). Under such conditions, computer-aided diagnostic systems [30, 36, 55] that segment these structures based on their color often fail.

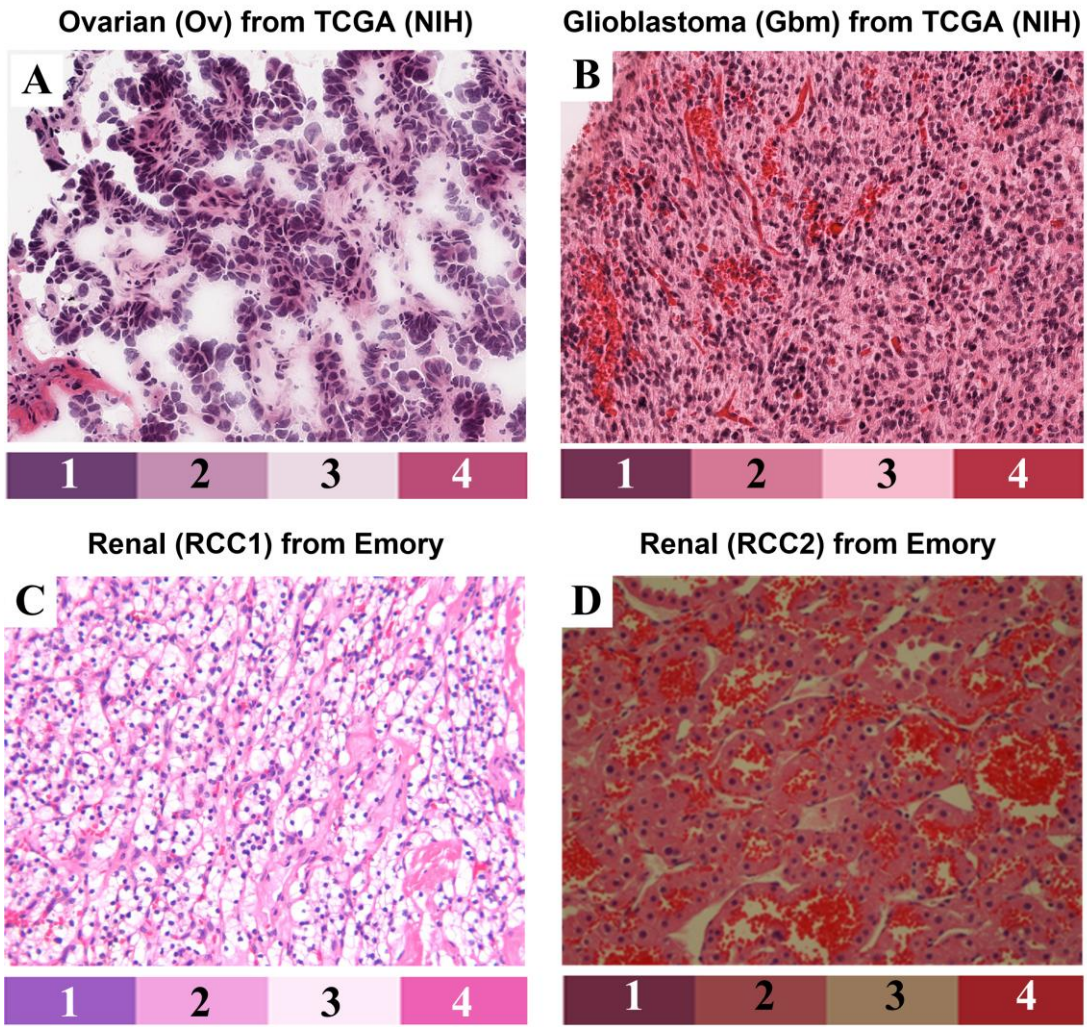

Figure 13: Color batch-effect between histopathological images in four datasets. A) ovarian (Ov), B) glioblastoma (Gbm), C) renal tumor (RCC1), and D) renal tumor (RCC2). Color palette illustrates cluster means of four H&E color classes. © 2011 IEEE.

One way to account for the observed difference in colors among images, i.e. 'batch effect', is to develop an interactive system that allows users to lend their domain knowledge to guide the segmentation process [36, 115]. However, user-interaction lowers the overall objectivity, reproducibility and speed of such systems. Among automatic segmentation methods, supervised learning techniques have been reported to be more accurate than unsupervised learning methods [41-43]. We find that these previous techniques are vulnerable to batch effect, and that they tend to perform well only for data from the batch on which they are trained (Table 5). Therefore, we propose a system for automatic color segmentation of histological images which is designed to be resistant to batch effect (Figure 14).

Our system incorporates knowledge from pre-segmented reference images to normalize (Figure 14, Step 1) and segment (Figure 14, Step 2) new patient images. Also, in order to make our system robust to the choice of reference image (*j*), we segment new images (*k*) with multiple reference images and combine labels, $\mathbf{L}_{0,j}^{k}$, using a voting scheme. Voting produces preliminary segmentation labels, $\mathbf{L}_{1}^{k}$, which we then use to reclassify (Figure 14, Step 3) test image pixels in their original color space and produce final segmentation labels, $\mathbf{L}_{2}^{k}$. The proposed system provides an automatic color segmentation of histopathological specimens that is resistant to batch effects. We achieve this by incorporating knowledge from domain experts into a novel color normalization scheme.

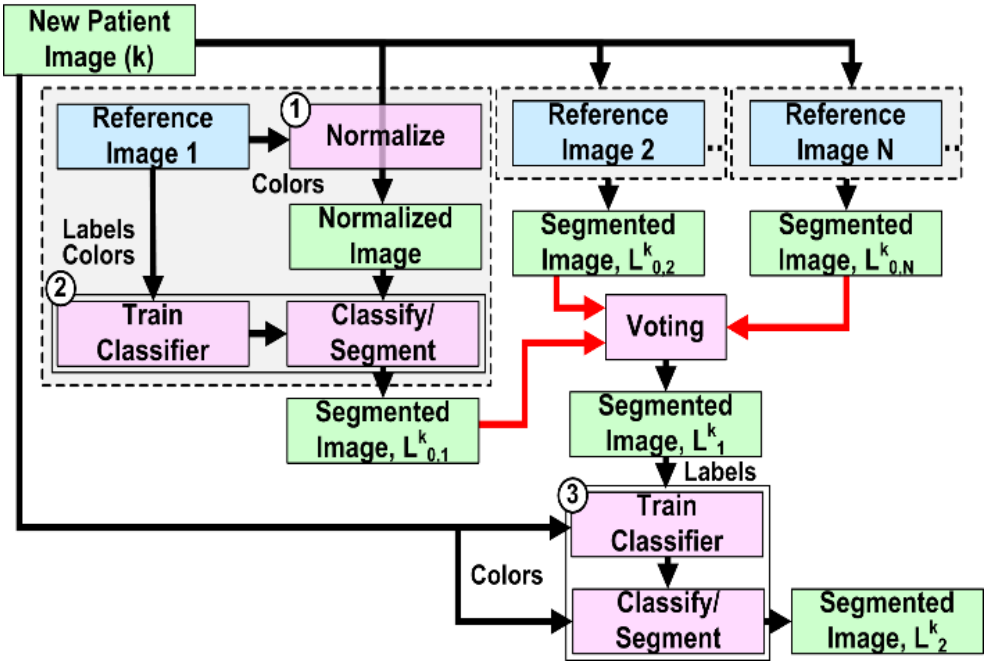

Figure 14: System flow diagram for batch-invariant, automatic stain segmentation. Three main steps: 1) normalize, 2) segment normalized image, and 3) re-classify pixels in the original color space. © 2011 IEEE.

Materials and Methods

Datasets

We analyze photomicrographs of H&E stained histological specimens. As we discussed in the chapter 1, H&E staining produces four distinguishable clusters of colors in the image—blue-purple (nuclear), white (no-stain or glands), pink (cytoplasm) and red (red blood cells). The color palettes in Figure 13 illustrate the mean color for each of the four color clusters in the ground truth segmentation. We consider four datasets in this chapter (Figure 13): two renal tumor (RCC1 and RCC2 with 55 and 47 images, respectively), one glioblastoma (Gbm, 52 images), and one ovarian (OvCa, 50 images). RCC1 and RCC2 were obtained at Emory University in separate experimental setups.

OvCa and Gbm images were obtained from TCGA repository [3]. To establish the ground truth labeling for each image, we developed an interface to help users label pixels semi-automatically. We use these labels to prepare reference images and to assess performance.

Image Normalization

We begin segmenting sample images by first normalizing the sample image's colors to the reference image's colors. Many color normalization techniques have been proposed [21, 22], including histogram or quantile normalization in which the distributions of the three color channels are normalized separately. Here, we mathematically describe quantile normalization of all pixels in an image. An image k contains N_k pixels where each pixel n is represented as a triplet given by

$$I^{k,n} = [R^{k,n}, G^{k,n}, B^{k,n}], \tag{9}$$

where $R^{k,n}, G^{k,n}, B^{k,n}$ are color channel intensity values. We define a rank function $\mathrm{f}_C^k \in \Re^{N_k \times N_k}$ that maps the color channel intensity, $C \in [R, G, B]$, from image k to a rank that ranges from 0 to N_k-1. Using the green channel as an example, $\mathrm{f}_G^k(\boldsymbol{G}^k) = \boldsymbol{r}_G^k$, where $\boldsymbol{G}^k, \boldsymbol{r}_G^k \in \Re^{N_k}$ are vectors of the green component intensity and rank for the k^{th} image, respectively. If $G^{k,n} \in [0,255]$ and $r_G^{k,n} \in [0, N_k - 1]$ are green component intensity and rank for the n^{th} pixel in the k^{th} image, then for any two pixels n_1 and n_2, $r_G^{k,n_1} \le r_G^{k,n_2}$ *iff* $G^{k,n_1} \le G^{k,n_2}$. The normalized green channel intensity of the n^{th} pixel of the k^{th} sample image to the j^{th} reference image can be computed using

$$\tilde{G}_j^{k,n} = \mathrm{h}_G^j\left[\frac{r_G^{k,n}}{N_k} \times N_j + \frac{1}{2}\right], \tag{10}$$

where $\mathrm{h}_G^j\left(r_G^{j,n}\right) = G^{j,n}$ is the inverse of the j^{th} image's green rank function $\mathrm{f}_G^j(\boldsymbol{G}^j) = \boldsymbol{r}_G^j$.

Figure 15 illustrates the quantile normalization process for two sample Gaussian distributions: test (distribution 1) and reference (distribution 2). Based on the number of points that have a value less than a definite value, we determine a rank and value relationship for both distributions. Thereafter, we normalize values in the test distribution by assigning values in reference distribution with the same rank as the test distribution.

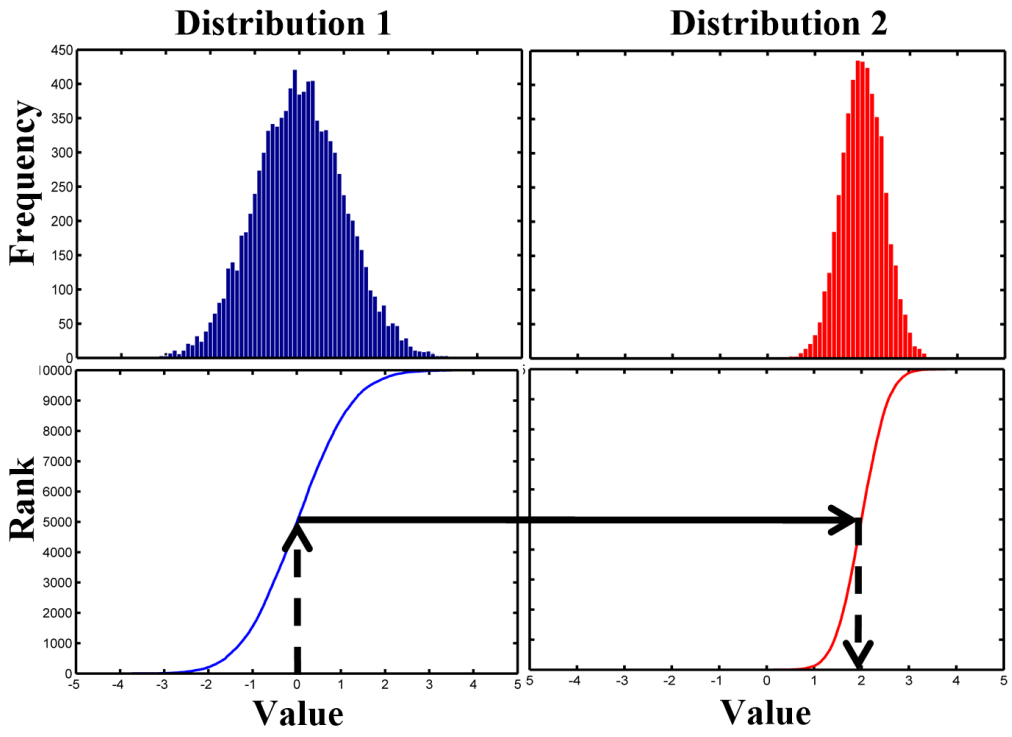

Figure 15: Quantile normalization of two sample Gaussian distributions.

We propose an alternative to simple quantile normalization, where we use the color map of the image instead of all pixels in the image. The color map is obtained by extracting the unique colors in the image. Therefore, compared to all pixels, the color map does not include the frequency of any colors. Mathematically, quantile normalization of color map elements is similar to that of all pixels except that the image is represented by a list of unique color triplets given by

$$U^{k,m} = [R^{k,m}, G^{k,m}, B^{k,m}], \tag{11}$$

where $m \in [0, M_k - 1]$ and M_k is the number of unique colors in the image. Because of variations in morphology from image to image, color and class frequencies vary. Figure 16 illustrates the distribution of green component intensity for all pixels and for color map elements of the four images in Figure 13. While the distributions of all pixels contain peaks which vary with changes in morphology and class prevalence, the distributions of color map shows less change between images. Therefore, normalizing the all pixels distributions rather than the color map distributions tends to distort colors in the normalized image. Once colors have been normalized, pixels are then classified by color.

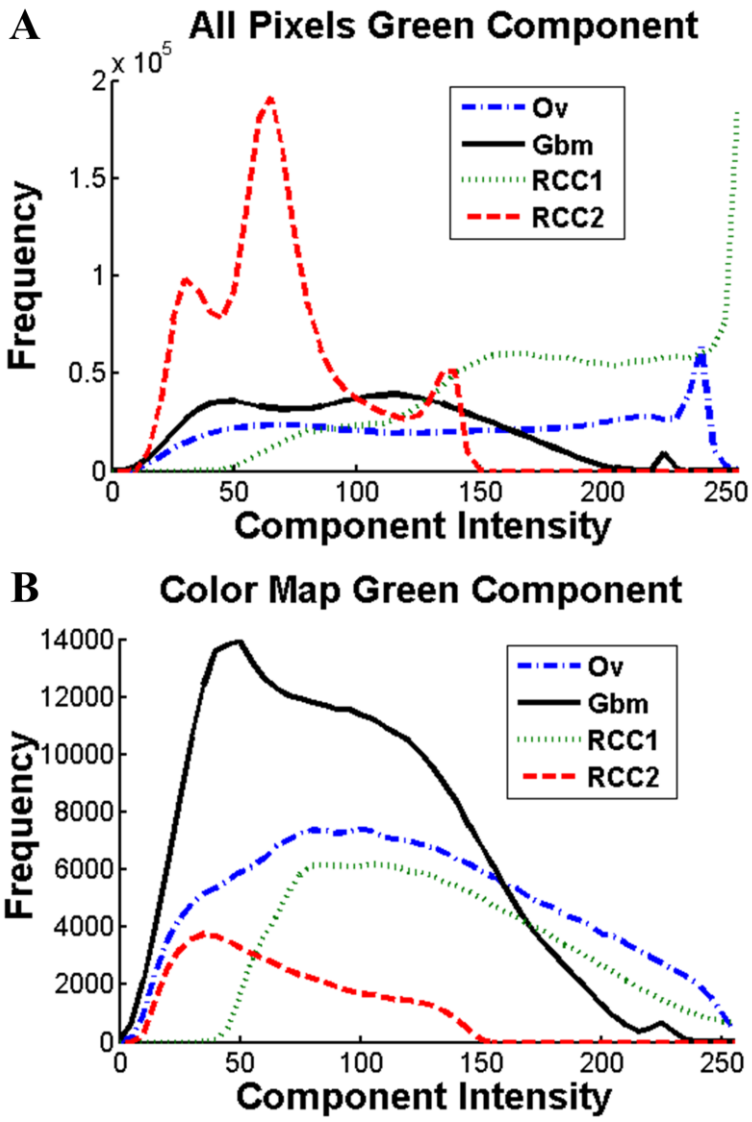

Figure 16: Distribution of green component intensities of (A) all pixels and (B) color map of the images in Figure 13.
Compared to color map, all pixels contain peaks which vary with changes in morphology and class prevalence. © 2011 IEEE.

Normalized Image Segmentation

Pixel classification is performed in the color space of a reference image. Using a four-class LDA classifier, we train using colors and labels obtained from ground truth segmentation of the reference image and classify pixels from the sample images based on normalized color. Let $\boldsymbol{L}^j \in \Re^{N_j}$ and $\boldsymbol{I}^j \in \Re^{N_j \times 3}$, where each element is $I^{j,n} = [R^{j,n}, G^{j,n}, B^{j,n}]$, be defined as the user-interactive segmentation labels and color values of pixels in image *j*, respectively. Let $\widetilde{\boldsymbol{I}}_j^k \in \Re^{N_k \times 3}$ be defined as image *k* normalized to image *j* where each element is given by

$$\widetilde{I}_j^{k,n} = [\widetilde{R}_j^{k,n}, \widetilde{G}_j^{k,n}, \widetilde{B}_j^{k,n}]. \tag{12}$$

For convenience, we define the function

$$\boldsymbol{L}' = \mathrm{LDA}\left(\boldsymbol{I}^j, \boldsymbol{L}^j, \boldsymbol{I}^k\right), \tag{13}$$

where $\boldsymbol{L}'$ contains segmentation labels for image $\boldsymbol{I}^k$ using an LDA classifier trained with pixel colors in $\boldsymbol{I}^j$ and labels in $\boldsymbol{L}^j$. $\boldsymbol{I}^k$ may also be a normalized image. Thus, to obtain the segmented image labels, $\mathbf{L}_0$ (Figure 14), we use

$$\boldsymbol{L}_{0,j}^k = \mathrm{LDA}\left(\boldsymbol{I}^j, \boldsymbol{L}^j, \widetilde{\boldsymbol{I}}_j^k\right). \tag{14}$$

The accuracy of segmentation depends on the choice of reference image. Therefore, in order to select optimal reference images, we perform CV within each dataset batch, where each image in the batch acts as a reference to normalize and segment all remaining images in the batch. We select the top 10 performing references from each batch. In order to avoid the choice of a single canonical reference image, in our system, a sample image is normalized and segmented 10 times, using a different reference image each time. For

each pixel in the sample image, we compute the final segmentation label by voting from multiple references. The label most frequently assigned to a pixel is chosen as its preliminary label (block $\boldsymbol{L}_1^k$ in Figure 14) before segmentation refinement.

Segmentation Refinement

The preliminary labels obtained by voting ($\boldsymbol{L}_1^k$, Figure 14) are good approximations of the ground truth labels, but we further refine this segmentation using the LDA classifier:

$$\mathbf{L}_2^k = \mathrm{LDA}\left(\mathbf{I}^k, \mathbf{L}_1^k, \mathbf{I}^k\right). \tag{15}$$

This step trains the LDA classifier using colors from the original sample image k and using labels from voting. The trained classifier is then used to re-classify all pixels in image k. Intuitively, this step ensures that the color groupings are separable in the original sample's image color space, and that any color distortion introduced by normalization is removed.

Results and Discussion

Comparison of Normalization Methods

Figure 17 compares color map and all pixels normalization of a test image (Figure 17.A) from RCC1 to a reference image (Figure 17.B) from RCC2. All pixel normalization forces the normalized image (Figure 17.G) to have more cytoplasm because the reference image is chromophobe subtype of RCC and has more cytoplasm

than test image, which has clear cell subtype. In contrary, color map normalized image (Figure 17.H) maintains test image morphology.

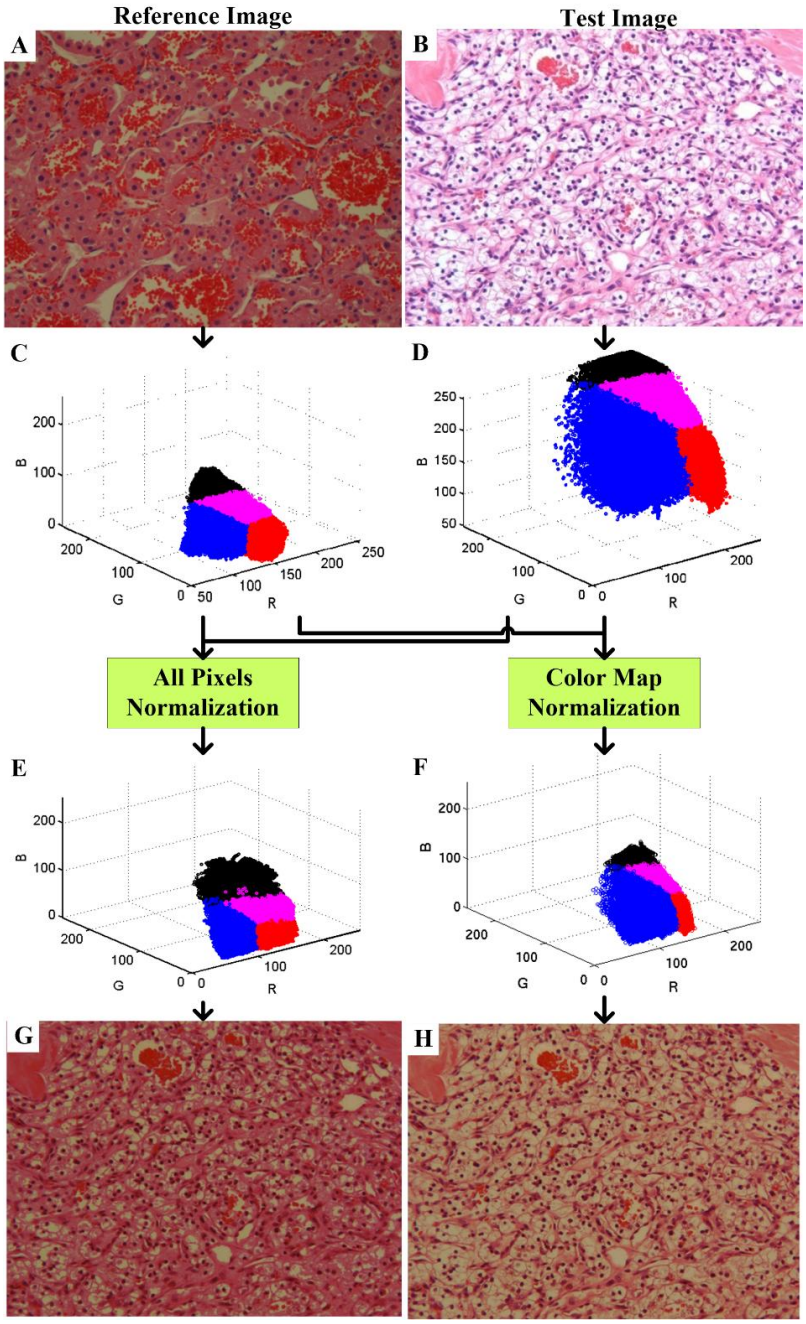

Figure 17: Results of color map and all pixel normalization of two renal tumor images from different batches.

Segmentation Performance With and Without Normalization

Table 5 lists the segmentation results from our system using two types of normalization (all pixels or color map) and compares them to our system with no normalization, i.e.

$$\mathbf{L}_{0,j}^{k} = \mathrm{LDA}\left(\mathbf{I}^{j}, \mathbf{L}^{j}, \mathbf{I}^{k}\right). \tag{16}$$

The overall performance, at 85%accuracy, is best for a system that uses color map normalization and re-classification (color map L2).

Table 5: Pixel-level, four-class segmentation accuracy for automatic stain segmentation system using color map, all pixels, or no normalization compared to ground truth.

Train	**Test**	**No norm.**		**All pixels**		**Color map**	
		L_1	**L_2**	**L_1**	**L_2**	**L_1**	**L_2**
RCC1	**RCC2**	0.17	0.08	0.83	0.82	0.79	0.82
	OvCa	0.32	0.43	0.79	0.83	0.77	0.81
	Gbm	0.22	0.54	0.82	0.85	0.80	0.82
RCC2	**RCC1**	0.37	0.56	0.85	0.88	0.79	0.87
	OvCa	0.32	0.40	0.82	0.85	0.86	0.87
	Gbm	0.23	0.46	0.80	0.84	0.81	0.84
OvCa	**RCC1**	0.13	0.13	0.73	0.78	0.82	0.87
	RCC2	0.16	0.16	0.72	0.74	0.85	0.84
	Gbm	0.77	0.82	0.76	0.80	0.85	0.85
Gbm	**RCC1**	0.13	0.13	0.82	0.84	0.85	0.87
	RCC2	0.16	0.16	0.78	0.80	0.84	0.83
	OvCa	0.84	0.84	0.78	0.83	0.87	0.87
Overall		**0.32**	**0.39**	**0.79**	**0.82**	**0.82**	**0.85**

* p-value for t-tests between: 1) L_2 all pixels and L_2 color map is—0.044, 2) L_1 and L_2 color map is—0.010.

Figure 18 compares segmentation results with color map and all pixels normalization. Re-classification (Figure 18, + and x) significantly improves the

segmentation performance. Color map normalization performs better than all pixels normalization except for four cases involving the RCC1 batch, possibly due to chromatic aberration, resulting in color map histogram distortion. However, in all pixels normalization, due to the low frequency of chromatic aberration colors, distortion is less severe.

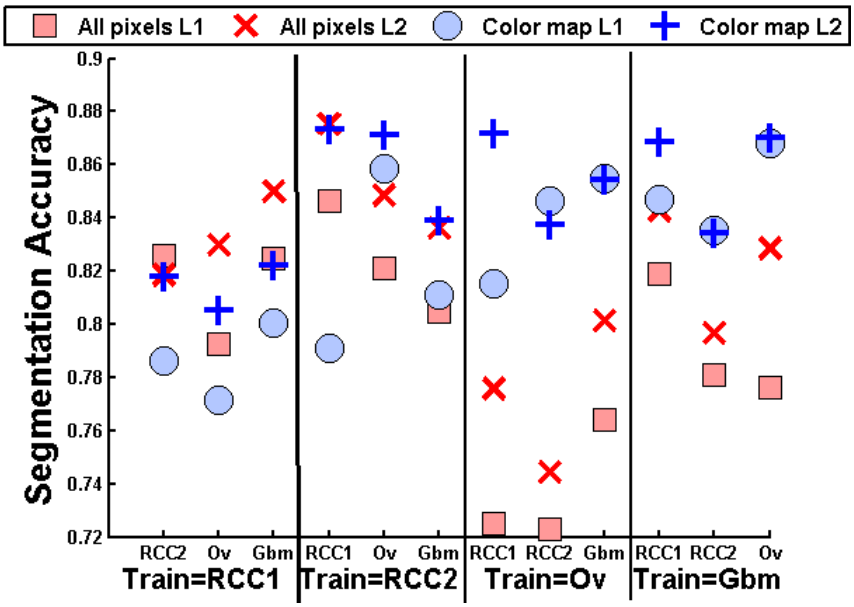

Figure 18: Comparison of segmentation accuracy of all pixels L_1, all pixels L_2, color map L_1, and color map L_2.

Figure 19 illustrates pseudo colored segmentation results for images in Figure 13.A and Figure 13.C. Figure 13.A is an OvCa batch image and is segmented on a system trained by reference images from the RCC2 batch. Figure 13.C is an RCC1 batch image and is segmented on a system trained by reference images from the Gbm batch. Again,

the re-classification step enhances the segmentation results and color map normalization retains the morphology of the test image. For instance, in Figure 19.G and Figure 19.H, all pixels normalization alters the morphology of the test image, over-segments the pink mask, and under-segments the white mask. Similarly, in Figure 19.B and Figure 19.C, the pink mask is over-segmented while the other three masks are under-segmented.

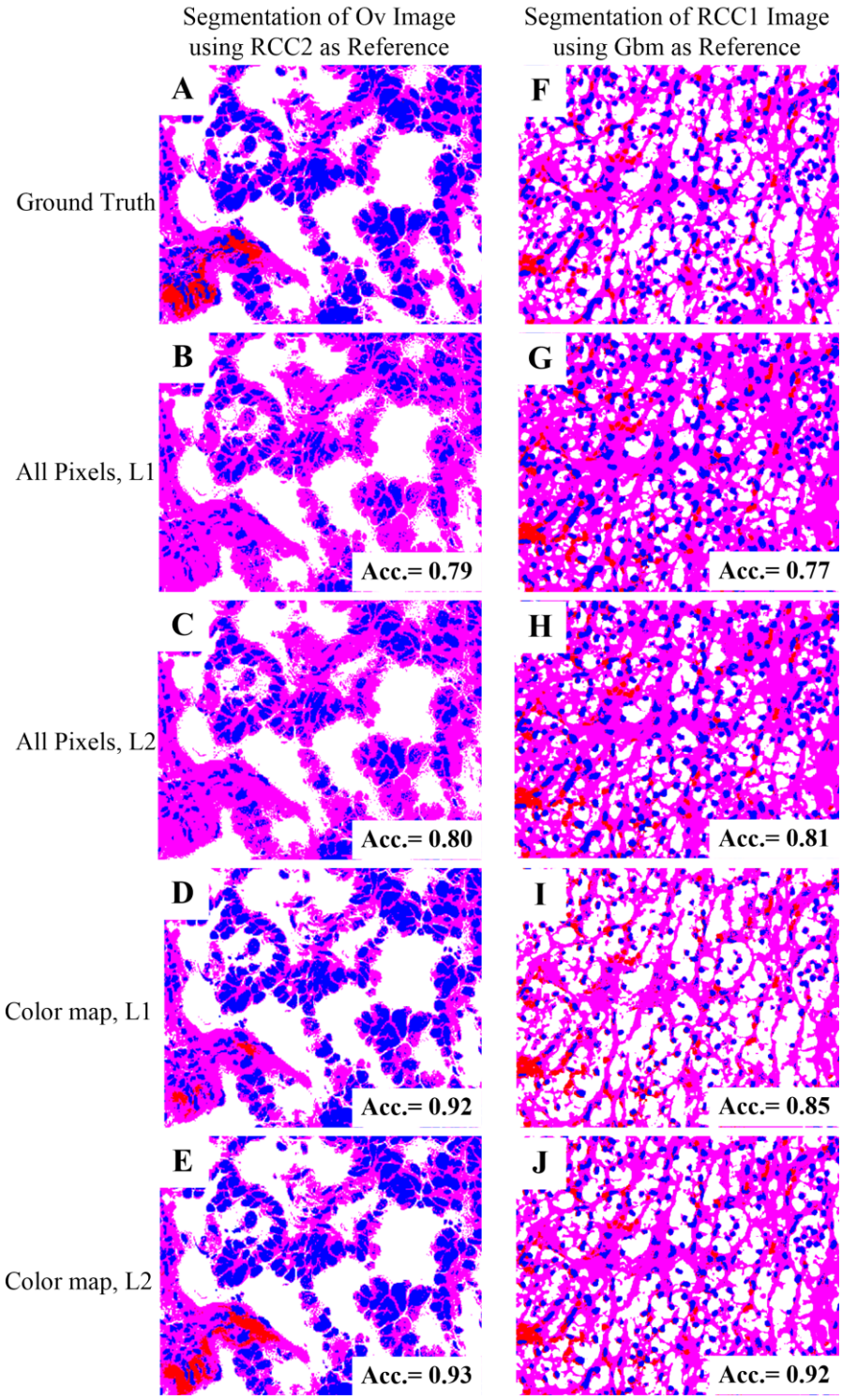

Figure 19: Segmentation of the images in Figure 13.A (top) and Figure 13.C (bottom). A magnified lower left portion of the image is shown; however, accuracy is reported for the full image. (A) Ground truth, (B) all pixels L1, (C) all pixels L2, (D) color map L1, and (E) color map L2. © 2011 IEEE.

Figure 20 illustrates an example segmentation scenario with three reference images, where re-classification resulted in a significant improvement in segmentation performance. First, a test image (Figure 20.D) was normalized and segmented using three references (Figure 20.A-C). As compared to ground truth (Figure 20.E), the classification accuracy for supervised segmentation using three references ranged between 0.72 to 0.86. After voting, the resulting L1 labels have 0.80 accuracy. Figure 20.I shows a 3-D scatter plot of original (un-normalized) RGB colors of the pixels, where each point is colored based on its L1 segmentation label. As marked by a green circle, the segmentation boundaries created by L1 labels are not smooth planes but are rather complex. This complexity is created because the voting method assigns the most frequent label to a pixel, irrespective of the label assigned to very similarly colored pixel. Thus, L1 labels are often incrorrect along the segmentation plane. The refine step re-classifies pixels using orginial colors and L1 labels as the training data. In this example, re-classification increased the segmentation accuracy to 0.89. Figure 20.J shows the 3-D scatter plot of original RGB colors of the pixels, where each point is colored based on its L2 segmentation label. This scatter plot has smoother segmentation planes and it is similar to the scatter plot of ground truth (Figure 20.F).

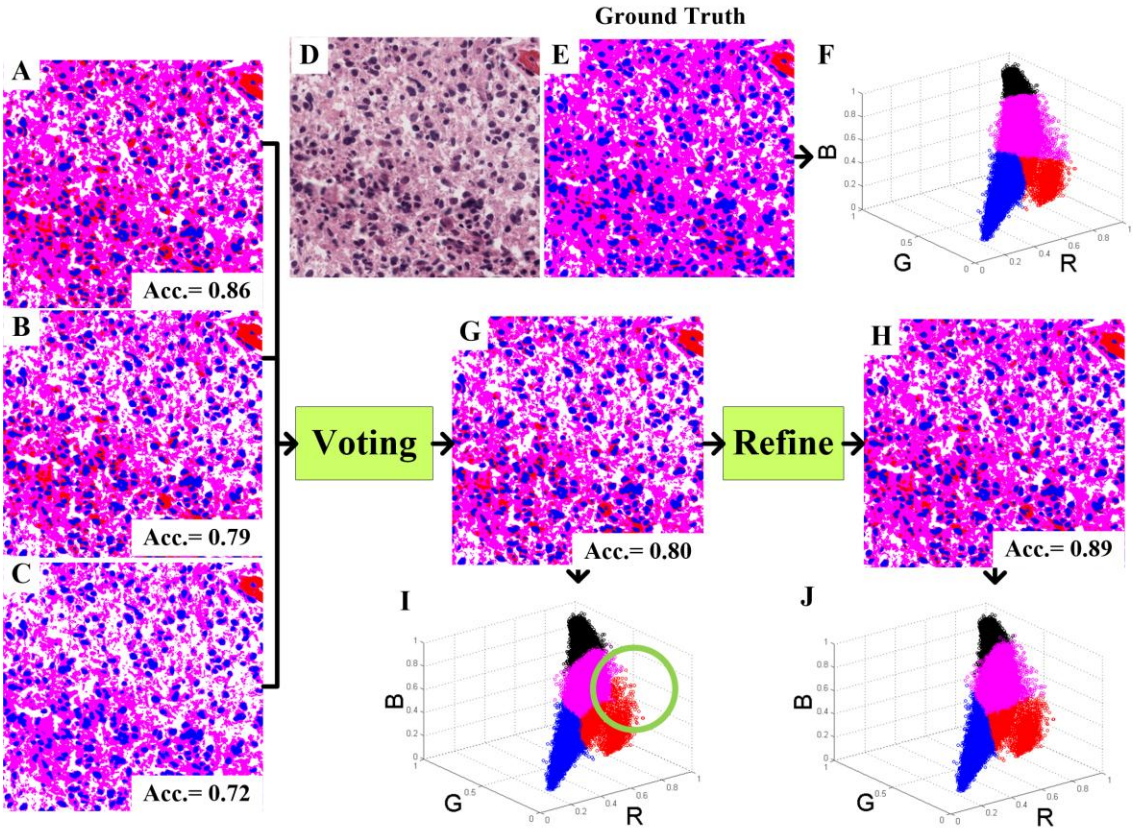

Figure 20: Effect of re-classification in original color space on segmentation performance.

Conclusion

In this chapter, we presented a novel supervised stain segmentation system for histopathological images that 1) incorporates domain knowledge to guide histological image segmentation and 2) normalizes images to reduce sensitivity to color batch effects. Results on four batches of H&E-stained histopathological images indicate that the performance of the supervised method is comparable to user-interactive expert segmentation. The high accuracy of stain segmentation masks will aid in increasing the overall performance and reproducibility of CDSSs. This supervised segmentation system can be easily trained and applied for the segmentation of microscopy images stained with other staining protocols.

CHAPTER 4

EDGE-BASED NUCLEAR CLUSTER SEGMENTATION

Introduction

Similar to Chapter 2 and 3, this chapter also focuses on developing a robust image segmentation method for histopathological images. The objective of the segmentation method developed in this chapter is to segment dense nuclear clusters in a binary nuclear-stain mask of a histopathological image. The research presented in this chapter was conducted in collaboration with other researchers and most of the content is part of published articles on nuclear segmentation [115, 116].

Pathologists often evaluate nuclear features such as nuclear count, nuclear shape and nuclear size to make important diagnostic decisions. In a healthy tissue sample, nuclei are mostly distinct and we can easily segment them using stain segmentation methods (such as the one discussed in Chapter 2). However, in a diseased sample, individual cells come close together and their nuclei form dense clusters. Therefore, it is necessary to segment these clusters before extracting nuclear features.

Researchers have proposed methods for the segmentation of simple-clusters and touching nuclei by extending and improving intensity- and morphological-segmentation methods [117, 118]. Few authors have developed algorithms that specifically address the challenge of nuclear cluster segmentation [46, 47, 119, 120]. Existing methods for nuclear segmentation have certain limitations: (1) some can only segment simple clusters [117, 118], (2) some make an assumption of circular nuclei, which is not always the case

especially in high-grade tumor images [46, 117, 118], (3) some result in nuclei with shapes that incorrectly depict naturally occurring nuclei [47, 119, 120], and (4) some are very complex and not easy to apply on large-scale datasets [46, 47, 119].

This chapter presents an edge-based image-segmentation method, which is simple to implement and can segment complex clusters with reasonable accuracy. The proposed method has the following five steps: (1) preprocessing to eliminate noise from nuclear mask and extract nuclear edges, (2) estimation of approximate nuclear area using shape-analysis on connected-objects in the cleaned nuclear mask, (3) detection of concavities on nuclear-cluster edges using the cross-product of adjacent tangents, (4) straight-line segmentation by connecting neighboring concavities, which result in regions larger than an area threshold, and (5) ellipse fitting on straight-line segmented regions. The elliptical model used is a good approximation to the original nuclear shape. Previously, Wang and Song [121] and Bai et al. [122] also developed methods for nuclear-cluster segmentation using concavity detection. Also, we perform a quantitative analysis of segmentation performance using simulated nuclear masks.

Materials and Methods

Preprocessing

Our approach is an edge-based method, so we preprocess the input RGB image of stained tissue sample to extract nuclear cluster edges. Figure 21 illustrates a renal tumor image after different preprocessing steps. First, we generate a binary nuclear mask from the RGB image using color segmentation (Chapter 3). The nuclear mask typically has

holes in the nuclear regions, noise at the edges and noise in the background (Figure 21.B). Second, we clean the nuclear mask using morphological processing. We fill the holes using morphological reconstruction [10]. We then eliminate the noise in the background using morphological opening and by removing the regions with area less than 20 pixels (Figure 21.C). Third, we smooth the cluster edges using a moving average low-pass filter (Figure 21.D). Smoothing eliminates noisy minor directional changes in cluster edges and preserves only true concavities. After preprocessing, we treat every nuclear cluster in the image separately.

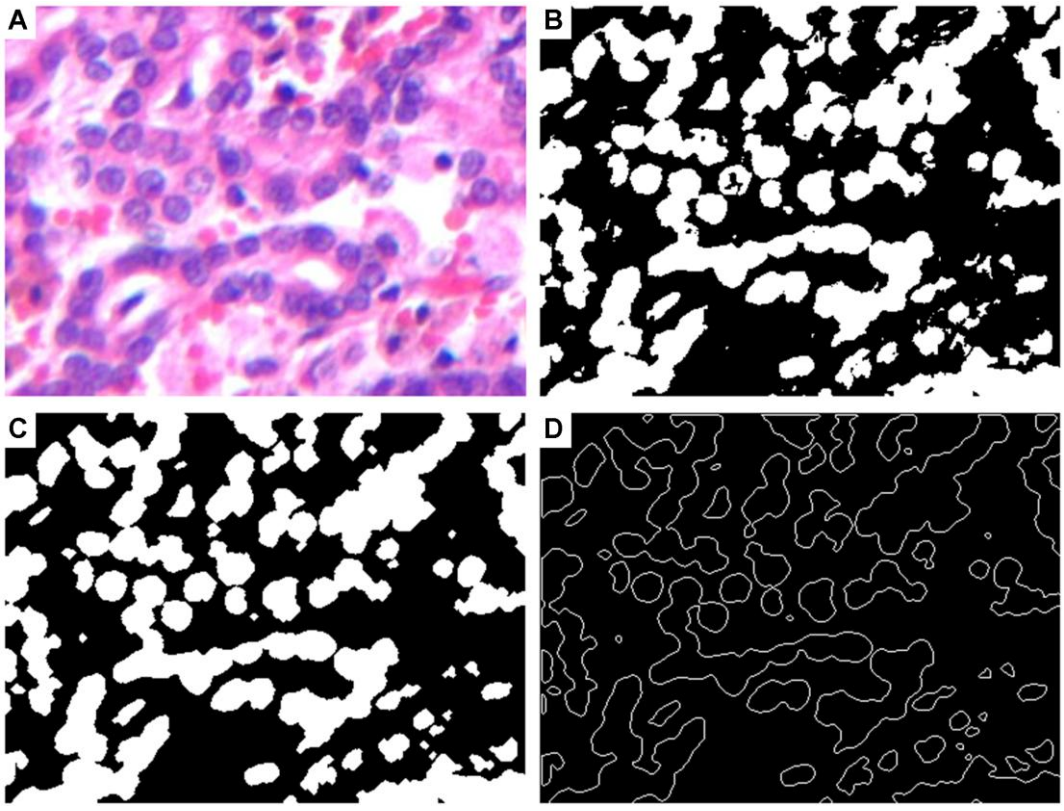

Figure 21: Preprocessing steps for nuclear segmentation on a renal tumor tissue sample (A) Input RGB image, (B) nuclear mask after color segmentation, (C) nuclear mask after morphological cleaning, and (D) cluster edges after edge smoothing. © 2009 IEEE

Approximate Nuclear-Area Estimation

Before segmenting the clusters, we estimate an approximate nuclear area (*A*) from the nuclear mask. We assume that the edge of a single nucleus is almost elliptical in shape while the edge of a nuclear cluster is not elliptical because it has multiple overlapping nuclei. Figure 22 illustrates the process of nuclear area estimation using elliptic deviation. First, we open the nuclear mask to separate touching nuclei and remove small noisy regions. Therefore, we select nuclear mask regions that have almost elliptical edges (elliptic deviation less than median value). We then estimate the median area of all the selected regions. Among the selected nuclear regions, we eliminate the regions with area less than the median value to avoid noise. We then set the value of *A* as the median area of all the remaining regions.

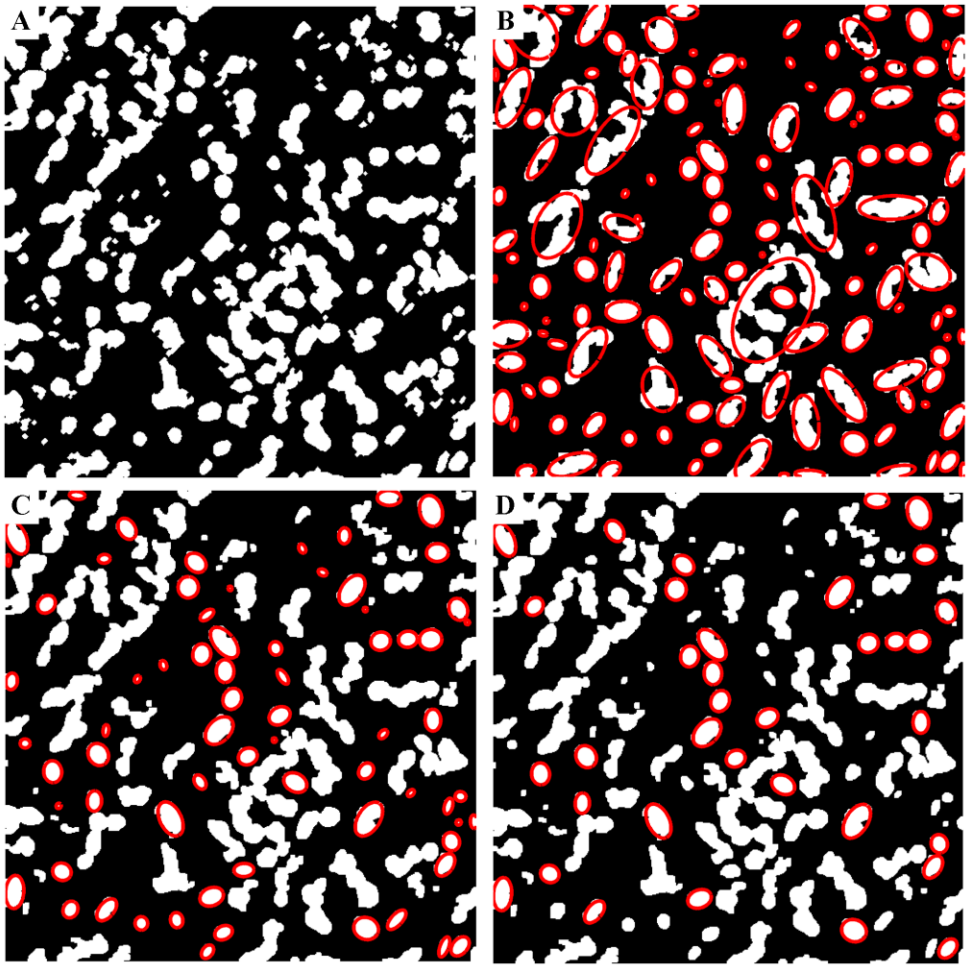

Figure 22: Estimation of nuclear area using elliptic deviation.
A Nuclear mask (A) for a renal papillary tumor image is morphologically opened to remove noise. (B-D) illustrate the process of selecting individual non-clustered nucleus samples. (B) Red curves show fitted ellipses for all regions in the opened mask. (C) Red curves show fitted ellipses for regions with elliptic deviation less than median value among all regions. (D) Red curves show fitted ellipses for regions with area more than median value among all regions in (C).

Concavity Detection

When two individual nuclei overlap, they form a notch or concavity at the points where the nuclear-cluster edges overlap. Therefore, we segment nuclear clusters at these points. We detect concavities using the cross product of adjacent tangential vectors while moving along a cluster edge in one direction. Figure 23 illustrates concavity detection for a cluster.

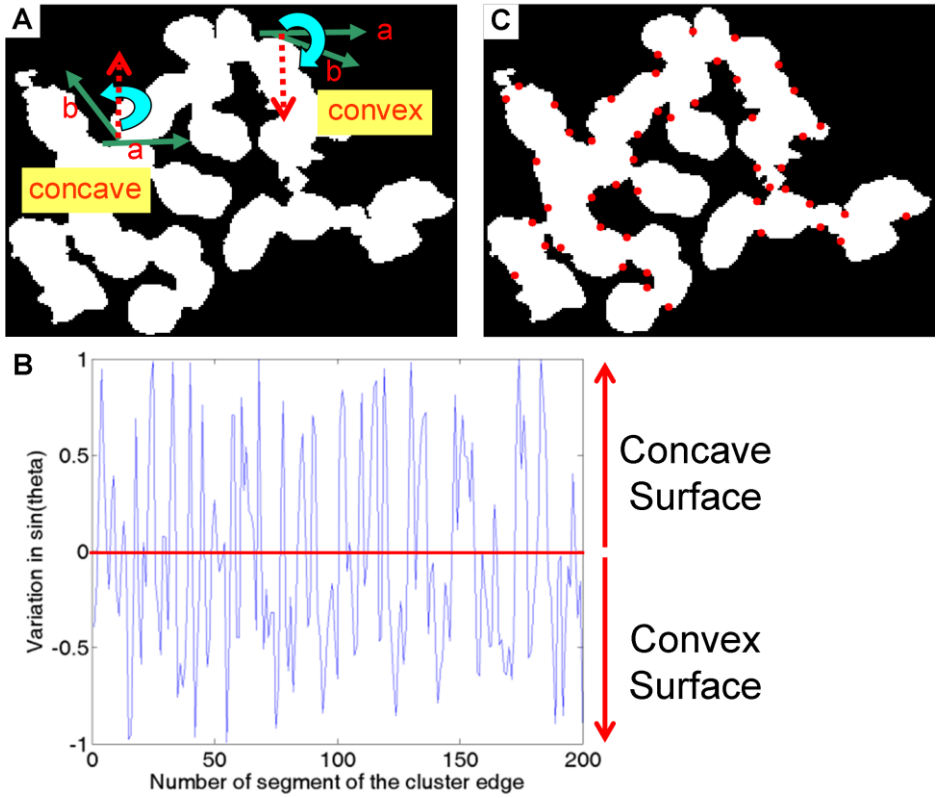

Figure 23: Concavity detection in a renal papillary cluster.
(A) A nuclear cluster with vectors *a* and *b* and the direction of sin(θ) marked at concave and convex edge points, (B) graph depicting variation in sin(θ) with segment number of the cluster edge, (C) concavities detected in the cluster. © 2009 IEEE

The following steps describe the algorithm for concavity detection in a cluster:

(1) Divide a cluster edge into piecewise segments. The size of segment affects the performance of concavity detection. Too large segment often miss a concavity while too small segment (e.g. one pixel) often results is false positives and makes detection process computationally complex. Thus, we decide the size of segment based on the length of cluster edge L, a minimum segment-length threshold l and a minimum number of segment threshold n. If $\frac{L}{200} > l$, we divide the edge into 200 equal length segments. Otherwise, if $\frac{L}{l} > n$, we divide the cluster edge into l length segments. Otherwise, we segment cluster into one pixel segments. We have used $l = 5$ and $n = 15$ in our implementation.

(2) Determine the tangential vectors for every segment using endpoints, given by

$$\mathbf{a} = [p_{2x} - p_{1x}; p_{2y} - p_{1y}; 0], \tag{17}$$

$$\mathbf{b} = [p_{3x} - p_{2x}; p_{3y} - p_{2y}; 0] \tag{18}$$

where **a** and **b** are two adjacent tangential vectors defined by three adjacent points on the cluster edge—p_1, p_2, and p_3.

(3) Determine the cross product between **a** and **b** vectors. Since the *z*-component is zero for the **a** and **b** vectors, the cross-product extends in the *z*-direction and $\sin(\theta)$ can be given by

$$\sin(\theta) = \frac{1}{|\boldsymbol{a}||\boldsymbol{b}|}[\boldsymbol{a}_x\boldsymbol{b}_y - \boldsymbol{a}_y\boldsymbol{b}_x] \tag{19}$$

Figure 23.B illustrates the variation of sin(θ) along the edge of the cluster in Figure 23.A.

(4) Determine location of concavity using a list of sin(θ) values. The value of sin(θ) is in the positive and negative *z*-direction for concave and convex portions of the edge, respectively (Figure 23.A). To avoid multiple detections for a concavity, we select the local maximum peak among the peaks that are located within a squared distance of 0.01**A*.

Straight-Line Segmentation

Straight line segmentation is the first step in the segmentation of clusters. We calculate the distance between all concavities for a cluster and connect the concavities starting with the ones closest to each other. The concavities are connected only if the following conditions are met

(1) Large portion of the connecting line segment lies inside the cluster

(2) The sizes of resulting segmented regions are larger than an area threshold

$$\text{Resulting nuclei area} > \text{threshold} \times \text{A}, \tag{20}$$

where the threshold is decided depending on the nuclear size variation in a tissue image. We obtained good results for different types of images using a value of 0.5. Figure 24 depicts the iterations of straight-line segmentation of a cluster in Figure 24.A

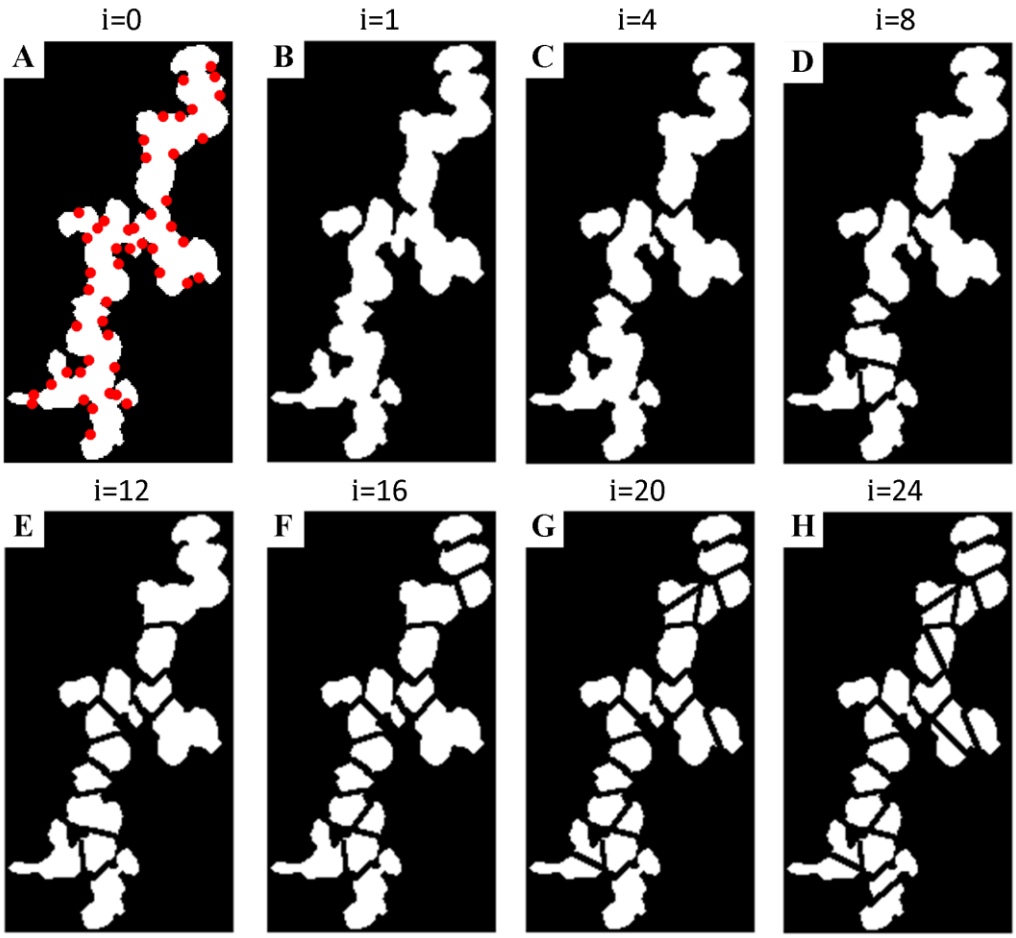

Figure 24: Iterations of a straight-line segmentation of a nuclear cluster using concavities.

Ellipse Fitting

Straight-line segmentation results in an approximate segmentation of nuclei in nuclear clusters. However, the resulting nuclei have sharp corners and their shape does not adequately model naturally occurring nuclei. Moreover, straight-line segmentation does not account for the overlap between adjacent nuclei. In ellipse fitting step, we take the knowledge from the straight-line segmented regions in the form of original-cluster-edge portion that belongs to an individual nucleus and model the nucleus using an elliptical model.

We apply the direct ellipse fitting method proposed by Fitzgibbon et al. due to its accuracy and simplicity of implementation [123]. The process of ellipse fitting includes the following steps:

(1) Sort straight-line segmented regions in a decreasing order of precedence depending on the amount of the original cluster edge it includes.

(2) Start ellipse fitting on the region with the highest precedence using the portion of straight-line-segmented region that is a part of the cluster edge. If the original cluster-edge length is less than 0.3 times of the complete edge of straight-line segmented region, original cluster-edge length lies on a straight line, or fitted ellipse is less than $0.5*A$, use complete edge of the straight-line-segmented region.

(3) Check the overlap between the present ellipse and previously fitted ellipses. If the overlap is less than 0.45 times the area of the fitted ellipse, fit the new ellipse, otherwise reject the ellipse.

(4) Repeat steps (2) and (3) till all the straight-line segmented regions are processed.

Once the ellipse fitting is completed for all clusters, we check if there is any portion of the mask which was not considered as a nucleus but is of sufficient size to be one. We provide the edges of such regions to the ellipse fitting algorithm. This step helps in reducing the number of missed detections. Figure 25.D depicts the iterations of ellipse fitting on the straight-line segmented cluster in Figure 25.A.

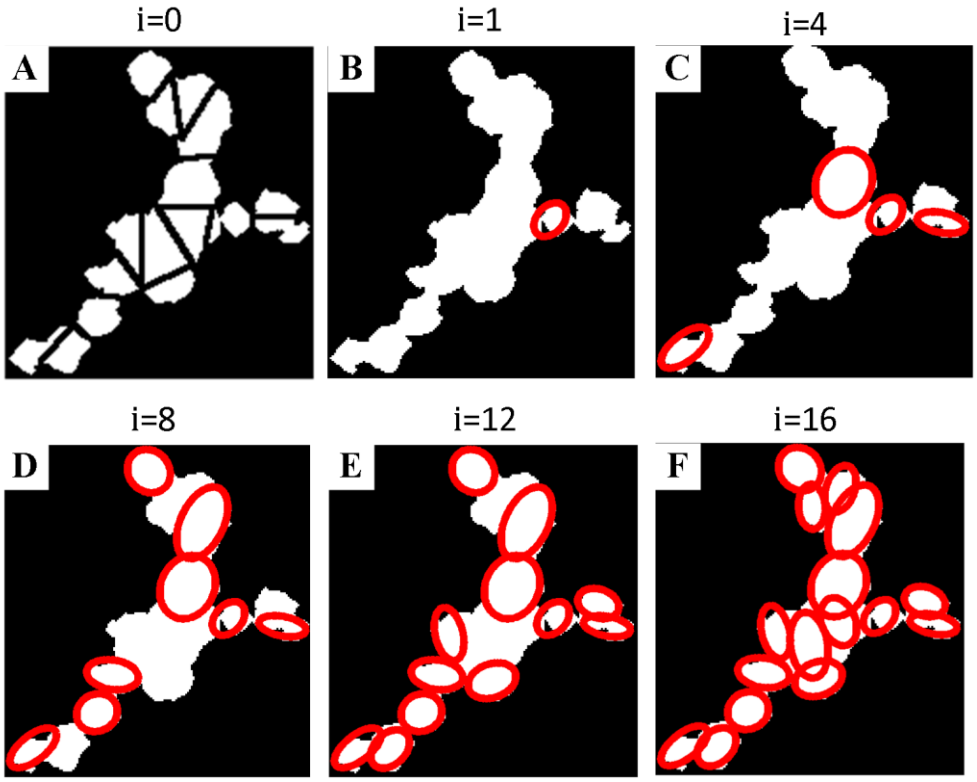

Figure 25: Iterations of ellipse fitting on a straight-line segmented cluster.

Simulated Data Generation

In this chapter, we use photomicrographs of H&E-stained biopsy tissue sections of renal, ovarian, and glioblastoma tumors. In addition to real tissue images, we also test the method on simulated nuclear mask images. In our simulations, we assume an elliptical model for nucleus given by following equations:

$$x = C_x + a\cos\theta\cos\emptyset - b\sin\theta\sin\emptyset \quad (21)$$

$$y = C_y + a\cos\theta\sin\emptyset + b\sin\theta\cos\emptyset \quad (22)$$

Where, as θ varies from 0 to 2π, x and y define the edge of a nucleus. a and b are semi-major and semi-minor axis lengths, ϕ is the inclination of the major axis w.r.t x-axis, and (C_x, C_y) is the center of the nucleus. For simulation, we select semi-major and semi-minor axis lengths using two bounded normal distributions. The means of the normal distribution for the semi-minor and the semi-major axis length are at 8 and 10 pixels, respectively. The standard deviation (σ) of both the distributions is equal and it is one of the parameters in the simulation. We bound both the distributions at mean ± σ.

We simulate a nuclear mask image, *BW*, using following algorithm. 1) Initialize *BW* to a blank 512x512 binary image. 2) Generate a nucleus i.e. a filled ellipse with following parameters— (C_x, C_y) at an uniformly random position in the range of *BW*, ϕ in the range 0 to 2π , and semi-axis lengths from the two bounded normal distributions. 3) Add some uniformly random noise to the generated nuclear edge, so that it is more similar to a real nucleus in histopathological data. 4) Add the nucleus to the BW image if the overlap in area with the existing nuclei in the BW image is less than or equal to 60% of its area. 5) Repeat last three steps till the nuclear count is *N*, another parameter in the

simulation Figure 26 illustrates some examples of simulated nuclear masks with different values of N and σ. We can observe that the complexity of nuclear clusters increases with increasing N and σ.

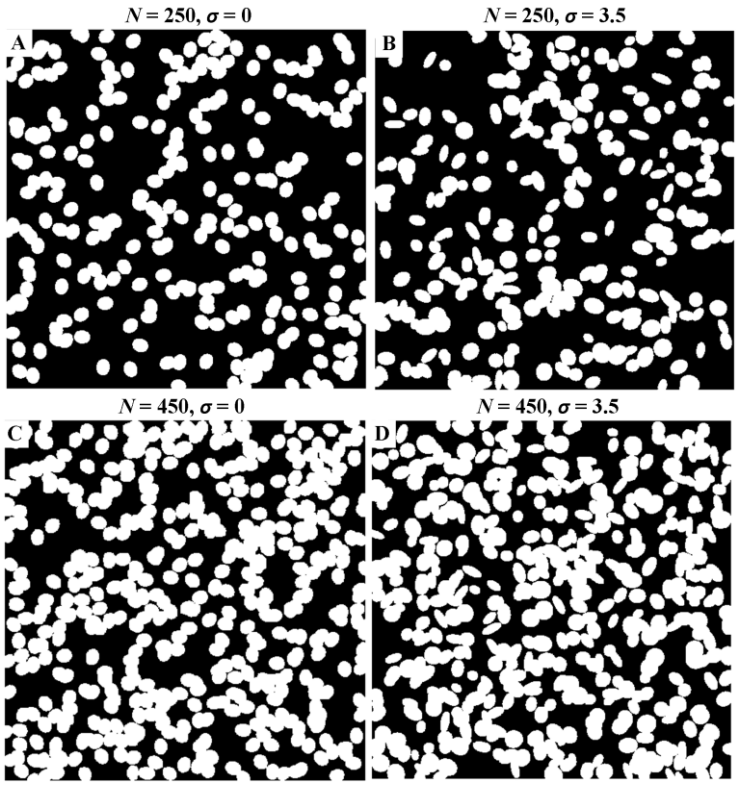

Figure 26: Examples of simulated nuclear mask with different parameters.
N: number of nuclei and σ: standard deviation in the axis lengths from their mean values.

Segmentation Evaluation

We evaluate the performance of our segmentation method by matching the predicted nuclear centers with the actual nuclear centers in the simulated data. While matching, we consider only mutually exclusive pairing of the centers. Let m and n are

centers in the list of predicted centers, *M*, and actual centers, *N*, respectively. Then *m* and *n* are a pair if they satisfy the follwing three conditions:

$$d(m,n) < d(m,k) \forall k \in N, k \neq n \quad (23)$$

$$d(m,n) < d(l,n) \forall l \in M, l \neq m \quad (24)$$

$$d(m,n) < 10 \pm \sigma \quad (25)$$

After the matching process, we calculate two performance metrics –TPR and FDR. TPR is the ratio of the number of matches to the total number of actual nuclei. FDR is the ratio of the number of unmatched predicted nuclei to the total number of predicted nuclei. High TPR and low FDR represent good performance of the method.

Results and Discussion

Performance of Nuclear Segmentation on Simulated Data

To visualize the performance of the method on the simulated data, we use a rectangular color-map visualization with various combinations of *N* and σ (Figure 27). We select *N* in the range of 100 to 450 with steps of 25, while we select σ in the range of 0 to 3.5 with steps of 0.5. Red color in the color map signifies high values while green color signifies low values. It can be observed that TPR > 0.8 (Figure 27.A) and FDR < 0.006 (Figure 27.B) for the majority of the simulation. Thus, the method is performing reasonably well.

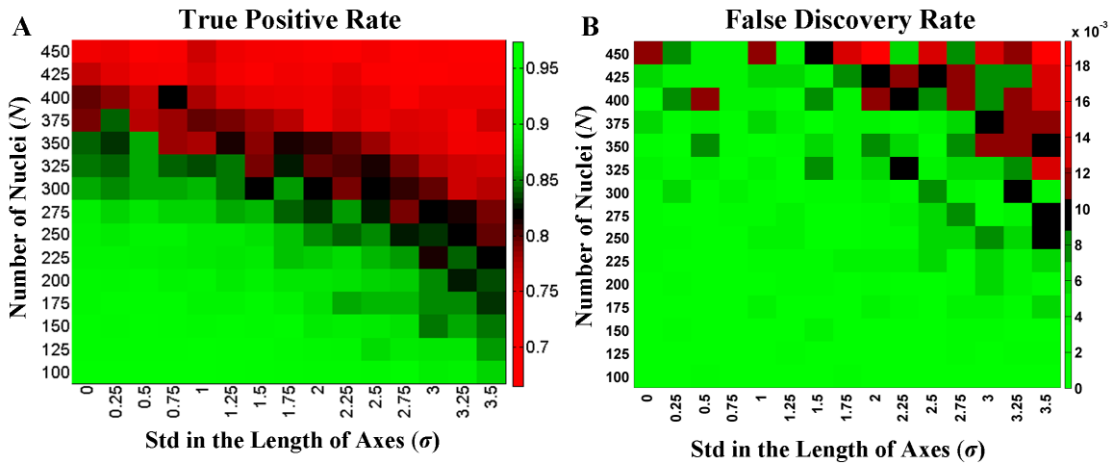

Figure 27: Performance of nuclear segmentation method on simulated data
(A) True positive rate is the ratio of number of matched predicted nuclei and total number of actual nuclei. B) False positive rate is the ratio of number of unmatched predicted nuclei and total number of predicted nuclei.

Performance of Nuclear Segmentation on Real Tissue Data

Figure 28 illustrates nuclear cluster segmentation results for tissue images of renal cell carcinoma (RCC), ovarian serous cystadenocarcinoma amd glioblastoma multiforme samples. It can be observed that segmentation method accurately segments nuclear clusters and single nuclei in the tissue images. It can also be observed that simulated nuclear clusters are similar to the clusters in the real tissue images. Therefore, the performance of the method on the simulated data is a good representative of its performance on real tissue images. Figure 28 illustrates concavity detection, straight-line segmentation, and ellipse fitting steps on some examples of real nuclear clusters.

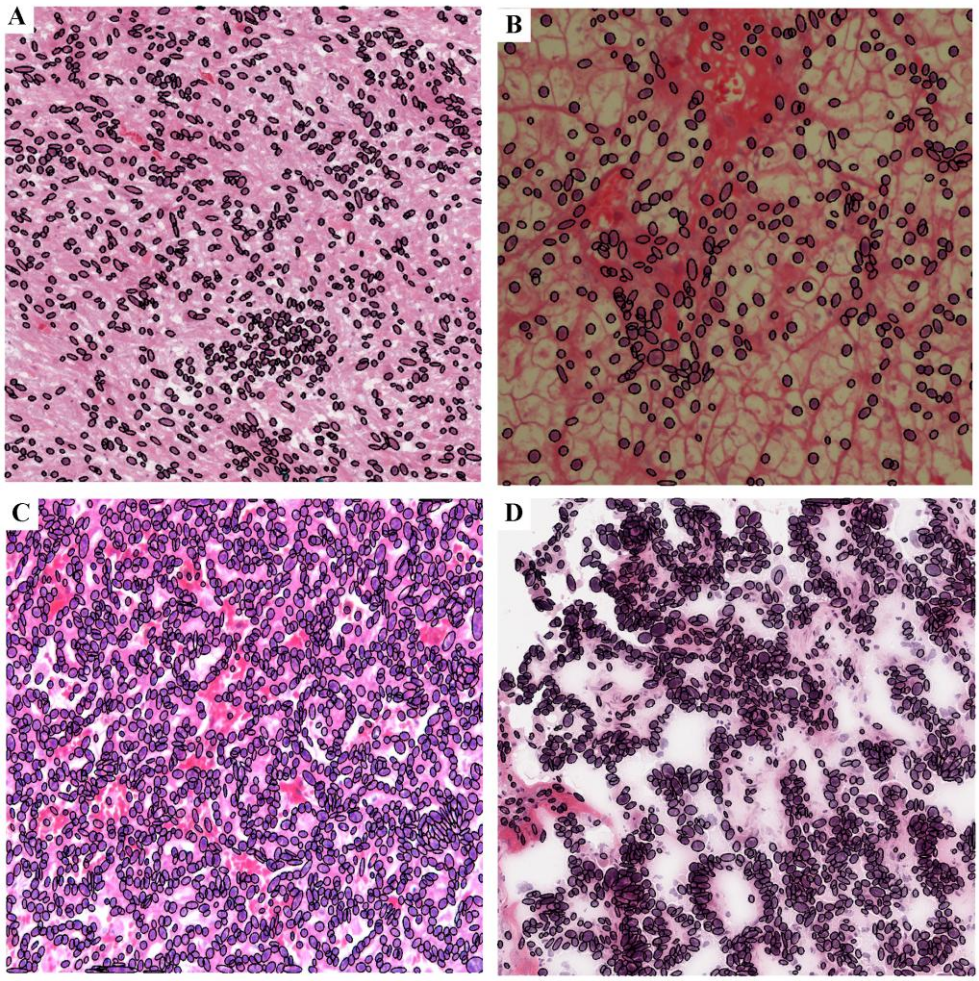

Figure 28: Nuclear segmentation results for real tissue images
(A) papillary RCC, (B) clear cell RCC, (C) ovarian serous cystadenocarcinoma, (D) glioblastoma multiforme.

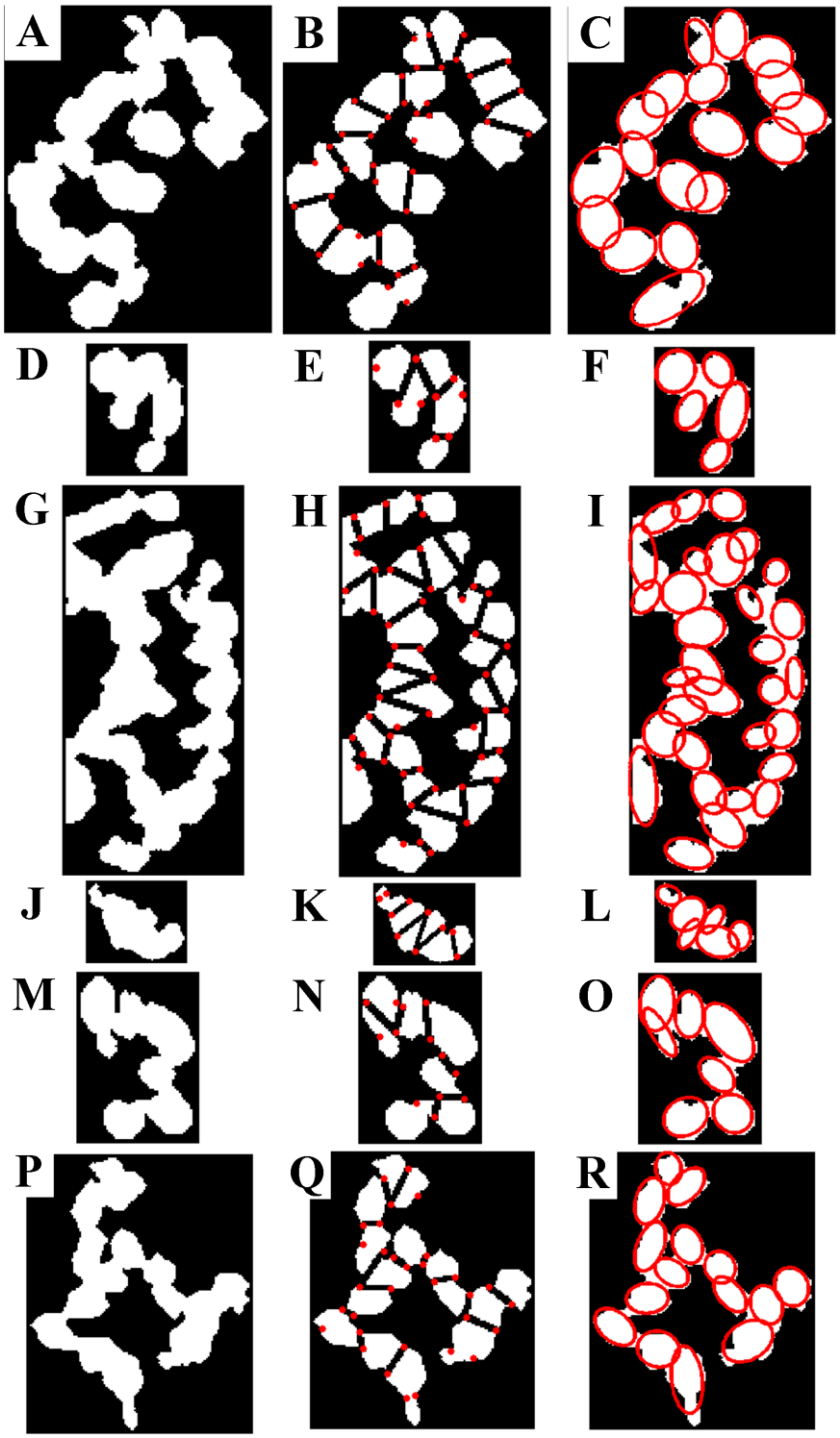

Figure 29: Examples of concavity detection, straight-line segmentation, ellipse fitting on real nuclear-clusters.

Conclusion

In this chapter, we presented a novel nuclear cluster segmentation method, based on concavity detection and ellipse fitting, for further segmenting dense nuclear clusters in the nuclear stain mask. We quantitatively evaluated the performance of the method on simulated data and visually on H&E-stained histopathological images. High performance on simulated data indicates that this method will be useful in segmenting elliptical shapes from complex structures of overlapping ellipses. Nuclear cluster segmentation is essential for extracting accurate nuclear-based features, which have proved to be useful for various cancer endpoints. We will apply this method for extracting informative nuclear features from renal tumor images.

CHAPTER 5

COMPREHENSIVE DESCRIPTION OF HISTOPATHOLOGICAL IMAGES

Introduction

This chapter addresses the informatics challenge of reducing semantic gap by developing a comprehensive set of image features and illustrating biological interpretation of emergent features. The research presented in this chapter was conducted in collaboration with other researchers and most of the content is part of a published article [124].

In the last few decades, many diagnostic models have been developed for cancer histological images to objectively and quickly predict disease endpoints (e.g., cancer grade or subtype) [5]. As a result, the literature is rich with image feature extraction methods for histological images. These features capture one or more of the following image properties: color [30], texture [37-39], topology [125], and shape [50]. Previous work suggests that some feature extraction methods may work better than others for a specific endpoint. For example, even though both fractal and multiwavelet methods all capture texture properties, Jafari-Khouzani and Soltanian-Zadeh [38] suggest that models based on multiwavelet features are superior for Gleason grading while Huang and Lee [37] suggest that fractal features perform better. Moreover, Doyle et al. [125] suggest that topological features perform better than textural features for breast cancer grading. Therefore, because of varying image properties and multiple feature extraction methods that capture the same image properties, it is unclear which set of features is optimal for

new cancer endpoints. The focus of research in the field of computer-aided diagnosis using histology has been on developing new innovative feature sets for a specific cancer endpoint. However, little work has been done in developing a general model with a comprehensive list of features that can be applied to multiple endpoints [30].

In this chapter, we develop a comprehensive feature set that consists of 12 feature subsets. Each feature subset is a combination of different image features capturing specific image properties. Our goal is to evaluate the diagnostic performance of this comprehensive feature set when applied to a variety of disease endpoints. Diagnostic models based on these feature subsets vary in classification accuracy. Moreover, for each disease endpoint, specific feature subsets tend to emerge as part of the best-performing predictive models. Although the association of feature subsets with disease endpoints is data-driven, many of the associations can be interpreted biologically. This suggests that such a comprehensive analysis can reveal biological clues for disease diagnosis. We perform this study using 12 binary disease endpoints including 6 renal tumor subtype endpoints and 6 renal cancer grade endpoints.

Materials and Methods

Datasets

We perform this study on micrographs of H&E stained renal tumor tissue samples. In this chapter, we use two separately acquired datasets (Figure 30). Dataset 1 includes subtype information and contains 48 images with 12 images of each subtype—chromophobe (CH), clear cell (CC), papillary (PA), and oncocytoma (ON). Dataset 2

includes information for Fuhrman grade and subtype. Its 58 images include 20, 17, 16, and 5 images of CH, CC, PA, and ON subtypes respectively. Excluding benign ON samples, among the remaining 53 images, 13, 13, 13, and 14 images are of grade 1 to 4, respectively. For the subtyping study, we combine all samples from both of the datasets. For grading, we consider malignant tumor images from dataset 2. Each sample image is about 1200x1600 pixels.

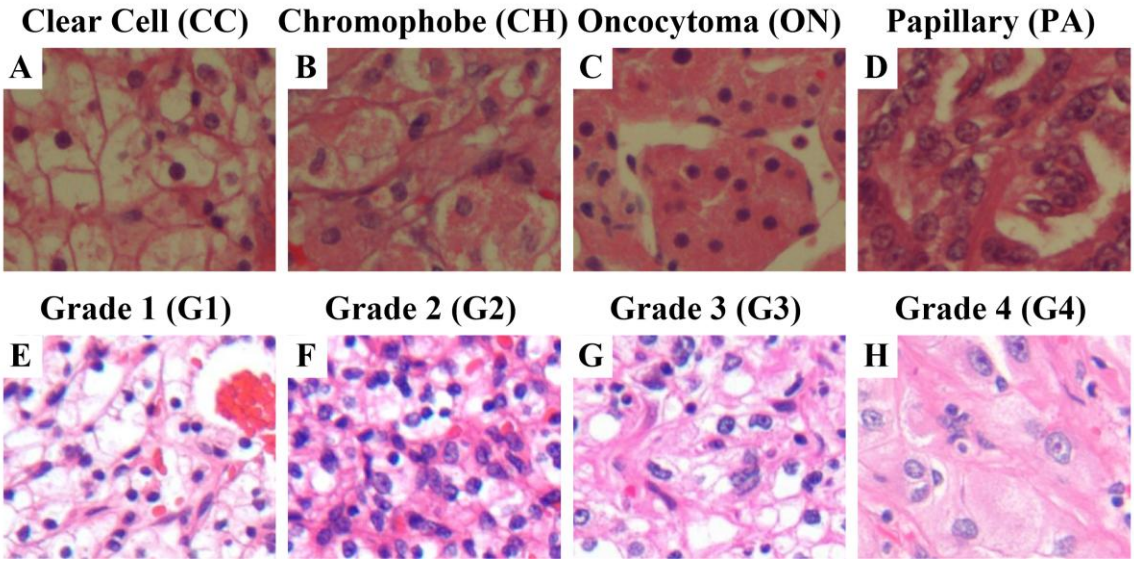

Figure 30: Sample histopathological tissue images for four renal tumor subtypes (A-D) from dataset 1 and four Fuhrman grades (E-F) from dataset 2.

Image Feature Extraction Methods

We crop 512x512 pixel non-overlapping, adjacent tiles from the central portion of each image sample. We extract features from each tile, and unless mentioned otherwise, we average features over all tiles to represent the sample. We extract a comprehensive set

of 2671 features from each sample. This set includes 12 feature subsets extracted from different processed forms of the original sample images. Table 6 lists the 12 feature subsets and their combination set (i.e., the ***All*** set).

Table 6: The comprehensive feature set including 2671 features.

Feature Subset	**Acronym**	**Count**
Color	C	48
Global Texture	GT	565
Global Color Texture	GCT	493
Stain Texture	StT	503
Cytoplasmic Stain Objects' shape	CStOS	249
Cytoplasmic Stain Texture	CStT	77
Glandular Objects' Shape	GOS	249
Nuclear Stain Objects' Shape	NStOS	249
Nuclear Stain Texture	NStT	77
Nuclear Stain Topology	NStTo	56
Nuclear Shape	NS	49
Nuclear Topology	NTo	56
All	A	2671

Figure 31 describes the flow of feature extraction, where green boxes represent different forms of the processed image while pink boxes represent feature subsets. We generate the "Normalized Sample Image" using a *color map* quantile normalization method (**Section 3.1**). For the "Color Quantized Image", we quantize the color space using SOM [39, 126] with the following parameters: 64 levels, 1-by-64 grid size, linear initialization along the greatest Eigen vector, and 'rectangular' lattice type. The "Stain Segmented Image" is a four-level grayscale image, where gray-levels of 1, 2, 3 and 4 correspond to nuclear, red-blood cells, cytoplasmic and glandular structures respectively. These structures correspond to distinct H&E color stains and we segment them using an automatic color segmentation method (Chapter 2). We then extract binary masks for "Nuclear", "Cytoplasmic" and "Glandular" structures in the image based on

segmentation labels. We further segment the nuclear clusters in the nuclear mask into individual nuclei to produce the "Segmented Nuclei" (Chapter 3).

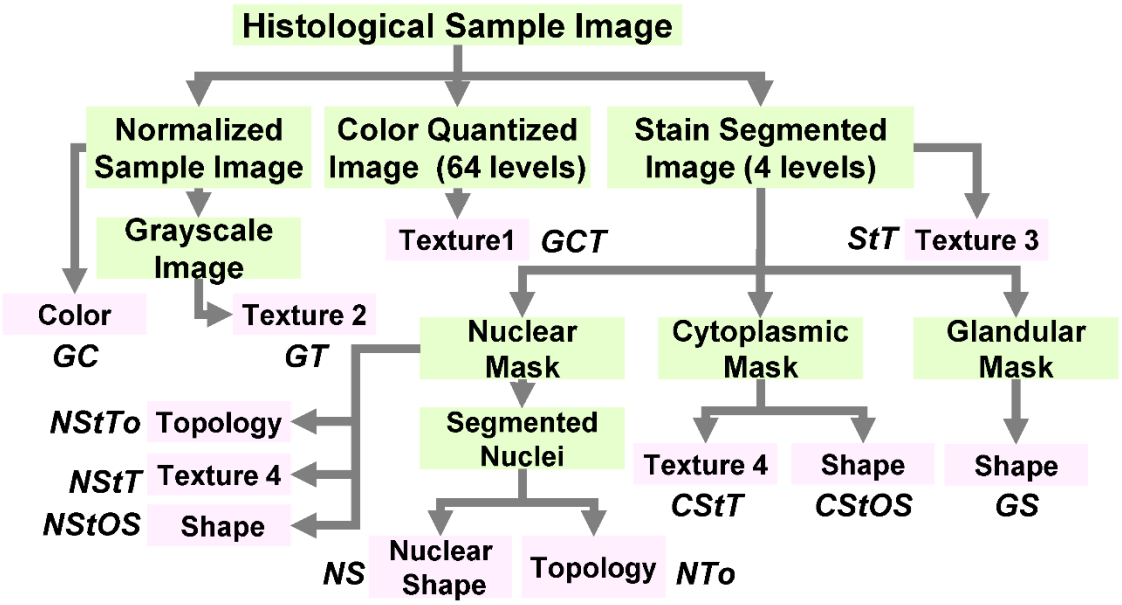

Figure 31: Flow diagram for image feature extraction. Green boxes: original or processed image. Pink boxes: feature subset.

The Color feature set correspond to distributions of R, G and B channel intensities with 16 bins per histogram [30]. We average each histogram bin over all tiles of an image to produce the Color feature subset.

The Texture1 feature set is a combination of Haralick, Gabor, wavelet packet and multiwavelet features. We extract a 64-level GLCM matrix for each tile and then we sum the GLCM matrices over all tiles. Thereafter, we extract 13 Haralick features from the summed GLCM matrix [127]. A 2-D Gabor filter is a Gaussian (with variances σ_x and σ_y

along x and y axes respectively) modulated by a sinusoid (with frequency f and orientation θ) [128, 129]. We consider σ_x =0.56/f and σ_y =2 σ_x. We obtain unique filters for different values of f and θ. We consider $\theta = \{0, \pi/4, \pi/2, 3\pi/4\}$ radians and $f \in \{\sqrt{2}/256, \sqrt{2}/128, \sqrt{2}/64, \sqrt{2}/32, \sqrt{2}/16, \sqrt{2}/8, \sqrt{2}/4\}$ cycles per pixel, therefore we have 28 distinct filters. For implementation, we consider a rectangular filter with range up to two standard deviations in both directions. We calculate energy (E) and entropy (H) [38] of each Gabor filter response image giving 56 features per tile. We average these features over all tiles to produce Gabor image features. We perform wavelet packet decomposition of the grayscale image using 'db6' and 'db20' wavelets [130]. We extract level-3 sub-matrices (total 64 sub-matrices per wavelet type) for each tile and then extract energy and entropy [38] of these sub-matrices. This results in 256 features per tile and we average over all tiles to produce wavelet packet image features. We also perform a two-level multiwavelet transform of the grayscale image with multiwavelets—GHM, SL and SA4 [38, 131]. We obtain 28 sub-matrices per multiwavelet type and calculate their energy and entropy resulting in 168 features per tile. The final multiwavelet features are an average of these 168 features over all tiles.

The Texture2 feature set is a combination of Texture1 features, gray-level distribution, and fractal features. Gray-level distribution captures the distribution of gray-levels in grayscale image using 64 bins in the histogram. Similar to color distribution, we average histogram bins over all tiles to produce 64 gray-level distribution features. We extract eight fractal dimensions for the grayscale image using the method described by Huang and Lee [37]. We calculate N_r and E_r for each tile and sum them to produce N_r and

E_r for the sample image. Thereafter, we calculate fractal dimensions from N_r and E_r using $s \in \{4,8,16,32,64.128,256\}$, where s^2 is the grid size.

The Texture3 feature set is a combination of Texture1 features and stain co-occurrence features. Stain co-occurrence captures the frequency of adjacent stains in a histological image [36]. We extract a 4x4 stain co-occurrence matrix similar to a GLCM matrix. This matrix is symmetric, so we extract 10 stain co-occurrence features from the lower triangular matrix. The final stain co-occurrence features for the image is the average of these features over all tiles.

The Texture4 feature set is a combination of gray-level distribution and Haralick features applied to specific stains. In the Texture1 feature set, we capture the grayscale texture of the image as a whole. In Texture4, we capture the grayscale texture of the cytoplasmic or nuclear stain areas in the grayscale image bounded by their respective masks.

The Shape feature set captures shape properties of the structures in three segmentation masks. Among the structures identified by segmentation, we eliminate the noise using a 20-pixel area threshold. The description for pixel area, convex hull area, solidity, perimeter, elliptical properties (area, major-minor axes lengths, eccentricity and orientation) and bending energy is available in [5, 50]. For ellipse fitting, we use the method described by Fitzgibbon et al. [123]. We extract boundary fractal dimension, using box counting on a binary object image. We extract Fourier shape descriptor error

(i.e., RMS error) in reproducing the shape using 1, 2, …, 20 harmonics [132]. We estimate the distribution of each of the 31 measures over the objects in all the tiles. We then represent this distribution using eight statistics: mean, median, minimum, maximum, standard deviation, inter-quartile range, skewness and kurtosis. In addition to these features, we also use object count as a feature. Therefore, the Shape feature set consists of 249 features (31*8+1).

The Topology feature set captures the spatial distribution pattern of objects in the histological image. We extract topology features using elliptical centers from unsegmented nuclear stain objects and segmented individual nuclei. We extract topology features by measuring properties of spatial graphs such Deluanay triangulation (areas and side lengths), Voronoi diagram (area, side length and perimeter) and minimum spanning tree side lengths [125]. We also measure object closeness, which is the average distance of an object to its five closest neighbors. We represent the distribution of these seven topology measures for a single image using the same eight statistics used for Shape features, resulting in 56 features.

The Nuclear Shape feature set is a combination of nucleus count, elliptical properties (the same as those of Shape features) and cluster size. Cluster size measures the number of nuclei in a cluster. Because it is a distribution, we estimate the same eight statistics, as object shape features.

Feature Selection and Classification

In this chapter, we consider all binary endpoints comparing pairs of classes. Both grading and subtyping datasets have 4 classes and 6 endpoints. We develop classification models for all combinations of binary endpoint (total of 12) and feature subset (total of 13, Table 6). We consider four classification methods (Bayesian, Logistic Regression (LR), k-Nearest Neighbors (k-NN) and Linear Support Vector Machine (SVM)) over a fixed set of parameters for each classifier. For Bayesian, we consider both pooled and un-pooled variance with spherical and diagonal variance matrices resulting in four Bayesian models. For k-NN, we consider ten *k* values from 1 to 10, resulting in 10 k-NN models. For SVM, we consider 28 cost values (0.1:0.1:0.9, 1:1:9, and 10:10:100 (start value : step : end value)), resulting in 28 SVM models. Logistic regression has no additional parameters. For each classifier model, we consider five feature selection techniques including t-test, Wilcoxon rank sum test, Significance Analysis of Microarrays (SAM) [133], and two types of mRMR: mRMR-d (difference) and mRMR-q (quotient) [134]. We consider 45 feature sizes ranging from 1 to 45. Thus, for each combination of endpoint and feature subset, we use cross validation to find optimal classification models from among 9,675 models.

We identify optimal classification models for each endpoint using stratified nested CV with 10 iterations and 5 folds in both the outer and inner CV. The inner CV is used for identifying optimal model parameters (i.e., feature selection method, feature size, classifier, and classifier parameters). The performance of each optimal model is then assessed using the testing set from outer CV. We select the simplest classification models

that are within one standard deviation of the best performing model. The simplest models are defined as those with the smallest feature size, highest k for k-NN models, and smallest cost for SVM models. For Bayesian models, we prefer pooled over un-pooled covariance and spherical over diagonal covariance. We have not assigned any preference to any particular classification method or feature selection method. Therefore, for each combination of endpoint and feature set, it is possible to obtain multiple optimal models. In such cases, we report the average performance of all models.

Results and Discussion

Classification Results

We optimize and validate models for every combination of feature subset and binary endpoints (both subtyping and grading endpoints). Figure 32 illustrates the scatter plots between optimizing CV accuracy and CV accuracy with the All feature subset. Each point in the scatter plot corresponds to average performance over 5-folds of one iteration in the outer CV loop and average performance over 250 iterations (5 outer folds*5 inner folds*10 inner iterations) in the inner optimizing loop. Most points are close to the diagonal line, suggesting that the performance of the inner optimizing CV predicts the performance of the outer CV. In the subtyping scatter plot, it can be observed that all but CH vs. CC and CH vs. ON perform with average accuracy > 90%. Low performance of these two endpoints is supported by the literature, as histologically and genetically, CH is similar to CC and ON. In the grading scatter plots, binary comparisons of grades differing by two or more levels tend to perform better (e.g., G1 vs. G4). Intuitively, this makes sense because with greater difference in grades, there are more visually apparent changes.

Table 7 and Table 8 list outer CV accuracy and standard deviation for subtyping and grading models using all feature subsets. The best performing subset for each endpoint is highlighted in red. It is interesting that the All subset is never the best performing subset. The gap between the best performing feature set and the All set is larger for grading endpoints. This is probably because, with the large feature list and fewer samples in the grading dataset, it is more likely that a model over fits. Hence, it is important to identify statistically important feature sets for an endpoint. Based on predictive performance, we can conclude that color, gray texture, nuclear stain object shape, nuclear shape, and topology are useful feature subsets for renal endpoints.

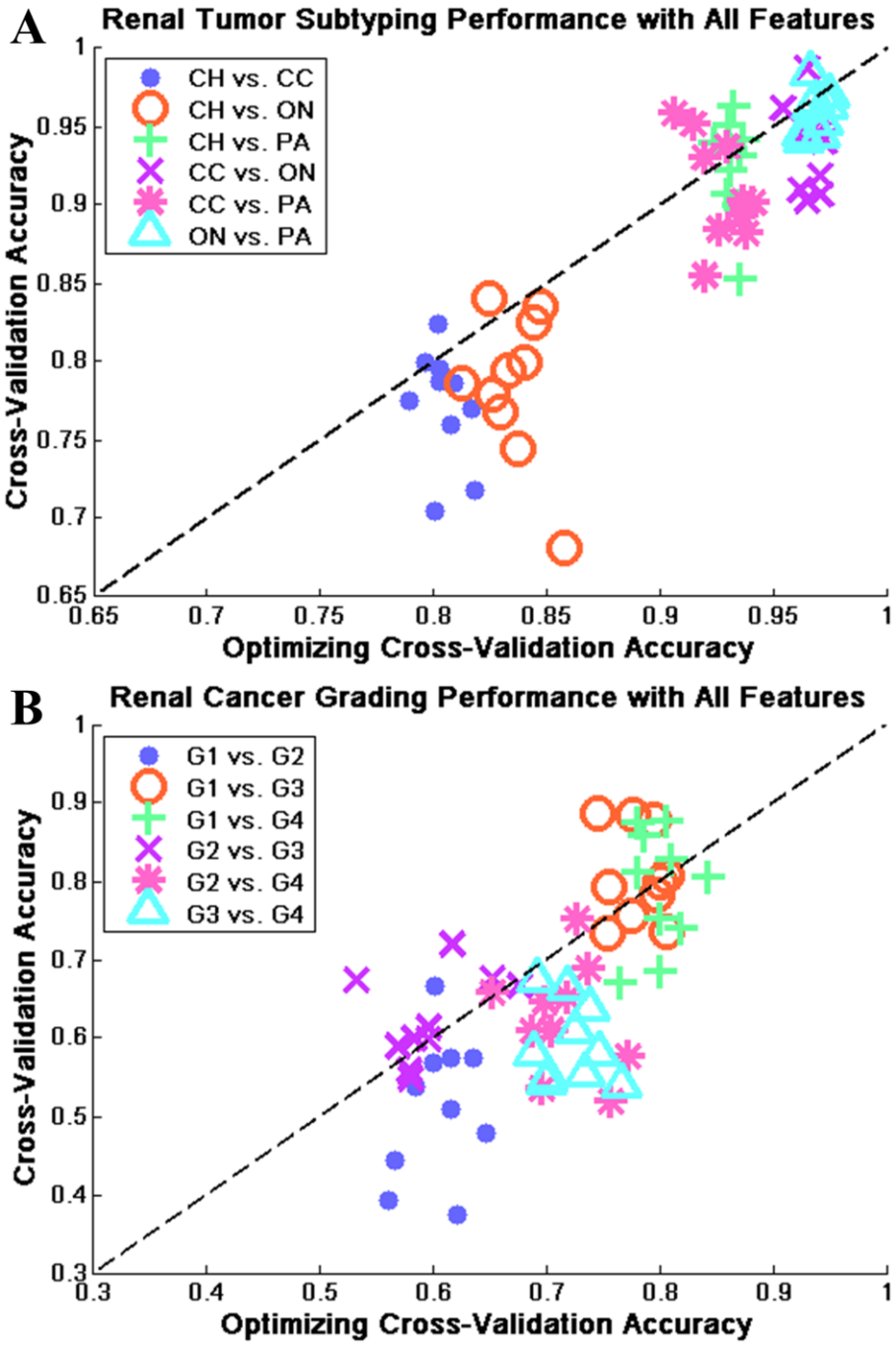

Figure 32: Scatter plot between optimizing CV accuracy and CV accuracy for binary renal subtyping and grading endpoints.
Each point in the plot represents an individual iteration of CV averaged over 5 folds.

Table 7: CV accuracy of binary renal subtyping models using different feature subsets in the comprehensive set.

Feature Subset	CH vs. CC	CH vs. ON	CH vs. PA	CC vs. ON	CC vs. PA	ON vs. PA
A	0.77±0.04	0.79±0.05	0.92±0.03	0.94±0.03	0.91±0.03	0.96±0.02
C	0.81±0.04	0.69±0.06	0.92±0.02	0.84±0.03	0.96±0.02	0.79±0.05
GT	0.84±0.03	0.62±0.03	0.89±0.02	0.92±0.02	0.91±0.03	0.77±0.03
GCT	0.65±0.04	0.65±0.02	0.84±0.02	0.65±0.06	0.86±0.05	0.84±0.04
StT	0.68±0.05	0.60±0.06	0.89±0.04	0.73±0.07	0.91±0.04	0.75±0.04
CStOS	0.77±0.04	0.63±0.02	0.48±0.04	0.60±0.07	0.66±0.02	0.58±0.03
CStT	0.73±0.05	0.64±0.04	0.86±0.04	0.70±0.05	0.80±0.03	0.81±0.06
GOS	0.69±0.03	0.72±0.06	0.70±0.04	0.92±0.04	0.88±0.04	0.74±0.05
NStOS	0.68±0.05	0.71±0.06	0.94±0.04	0.93±0.02	0.75±0.04	0.93±0.03
NStT	0.58±0.07	0.70±0.04	0.82±0.03	0.55±0.04	0.87±0.03	0.85±0.02
NStTo	0.73±0.03	0.76±0.05	0.60±0.06	0.58±0.05	0.56±0.05	0.74±0.07
NS	0.71±0.05	0.67±0.04	0.88±0.02	0.97±0.02	0.82±0.04	0.98±0.01
NTo	0.78±0.04	0.82±0.02	0.89±0.03	0.58±0.05	0.71±0.05	0.71±0.04

Note: Values in red best performing feature subset for any binary model.

Table 8: CV accuracy of binary renal grading models using different feature subsets in the comprehensive set.

Feature Subset	G1 vs. G2	G1 vs. G3	G1 vs. G4	G2 vs. G3	G2 vs. G4	G3 vs. G4
A	0.51±0.09	0.81±0.06	0.79±0.08	0.62±0.06	0.62±0.07	0.59±0.05
C	0.70±0.08	0.58±0.05	0.62±0.05	0.46±0.10	0.56±0.09	0.62±0.07
GT	0.47±0.04	0.51±0.09	0.74±0.08	0.43±0.08	0.63±0.07	0.66±0.13
GCT	0.52±0.06	0.54±0.10	0.65±0.07	0.49±0.04	0.55±0.10	0.51±0.08
StT	0.62±0.10	0.56±0.05	0.63±0.05	0.42±0.07	0.67±0.07	0.55±0.08
CStOS	0.44±0.06	0.48±0.06	0.59±0.09	0.52±0.09	0.68±0.09	0.63±0.06
CStT	0.48±0.10	0.43±0.08	0.62±0.06	0.50±0.07	0.54±0.04	0.59±0.06
GOS	0.44±0.08	0.55±0.06	0.49±0.08	0.44±0.07	0.44±0.08	0.44±0.06
NStOS	0.41±0.08	0.84±0.08	0.59±0.07	0.73±0.09	0.48±0.09	0.45±0.05
NStT	0.48±0.09	0.65±0.07	0.61±0.07	0.45±0.08	0.65±0.06	0.51±0.06
NStTo	0.37±0.09	0.45±0.10	0.49±0.08	0.45±0.07	0.47±0.07	0.47±0.08
NS	0.61±0.07	0.81±0.06	0.84±0.05	0.68±0.06	0.82±0.08	0.59±0.08
NTo	0.48±0.07	0.59±0.07	0.61±0.09	0.46±0.07	0.59±0.07	0.48±0.09

Note: Values in red best performing feature subset for any binary model.

Feature Ranking

In this section we will illustrate the important subsets that emerge for individual endpoints. First, for each endpoint, we consider all of the models that use the All feature set and calculate the percentage contribution of individual feature subsets. Figure 33 illustrates the average percentage contribution of feature subsets for both subtyping and grading endpoints. Among subtyping endpoints, all but CH vs. ON have a dominant contribution from one subset over the others (> 50%). These subsets are GT, NStOS, NS, C, and NStOS for CH vs. CC, CH vs. PA, CC vs. ON, CC vs. PA, and ON vs. PA, respectively. The same feature subsets are the best performing subsets in Table 7 for these endpoints. Among grading endpoints, only four have dominant contributing subsets. These subsets are NStOS, NS, NStOS and CStOS for G1 vs. G3, G1 vs. G4, G2 vs. G3, and G3 vs. G4, respectively. All but the G3 vs. G4 subset are the best performing subsets in Table 8.

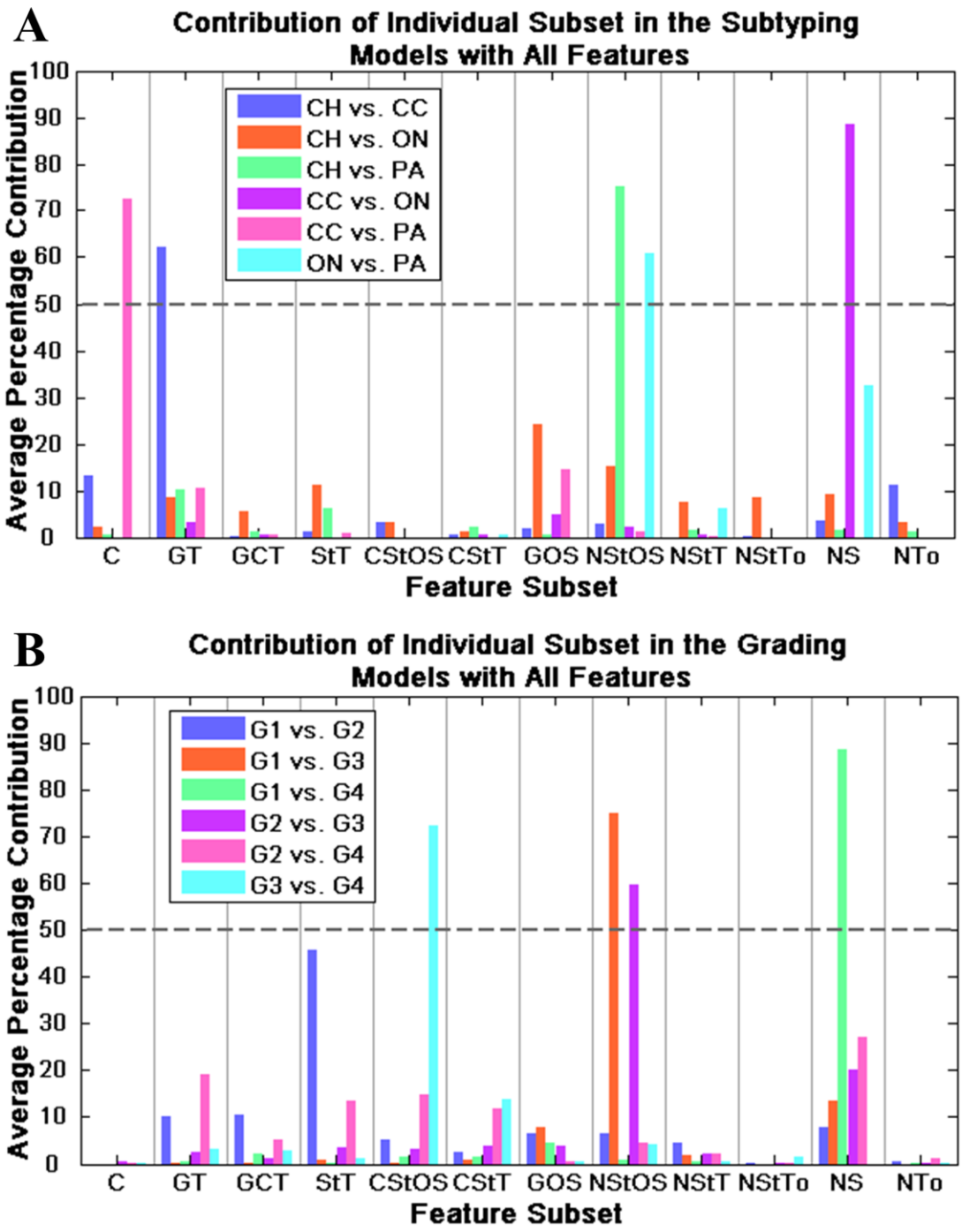

Figure 33: Contribution of feature subsets in the comprehensive set in renal tumor diagnostic models.
All subtyping endpoints have a distinct contribution (probability > 0.5) from a single feature subset with the exception of CH vs. ON. Similarly all grading endpoints have a distinct contribution from a single subset with the exception of G1 vs. G2 and G2vs. G4.

We further investigate the importance of feature subsets using a method normally used to identify over- represented Gene Ontology terms in a list of genes [135]. For each endpoint, we consider all optimal classification models that select features from the entire set of 2671 image features. We count the number of features drawn from each of the 12 feature subsets and use a one-sided Fisher's exact test to determine if any of the feature subsets are statistically over-represented at a p-value threshold of 0.01 (adjusted for multiplicity using the Bonferroni method). A small p-value for a subset indicates that the number of features selected from that subset is higher than what is expected by random chance. In Table 9, for each endpoint, we mark feature subsets that are statistically over-represented with X's. The results in Table 9 roughly correspond to those in Figure 33.

Table 9: Statistically over-represented feature subsets in diagnostic models for renal tumor subtyping and grading endpoints.

Feature Subset	**CH vs. CC**	**CH vs. ON**	**CH vs. PA**	**CC vs. ON**	**CC vs. PA**	**ON vs. PA**	**G1 vs. G2**	**G1 vs. G3**	**G1 vs. G4**	**G2 vs. G3**	**G2 vs. G4**	**G3 vs. G4**
C	X				X							
GT	X											
GCT												
StT							X					
CStOS											X	X
CStT											X	X
GOS		X			X							
NStOS		X	X			X		X		X		
NstT		X				X		X				
NStTo		X										
NS	X	X		X		X	X	X	X	X	X	
NTo	X											

Biological Interpretation

Eble et al. provide guidelines for subtyping renal tumors [136]. CC has clear cytoplasm with distinct cell membranes and round nuclei. CH has granular cytoplasm with prominent cell membranes and wrinkled nuclei. Perinuclear halos (i.e., white stain surrounding nuclei) are common in chromophobe images. PA has finger-like nuclear clusters. ON has granular cytoplasm with round nuclei arranged in compact nests or microcysts [136]. We can relate these biological properties to the feature subsets as follows. CH, with its wrinkled nuclei, granular cytoplasm and perinuclear halos, differs from other subtypes in nuclear, texture, and glandular object features. This may explain why the NS, NStOS, GT and GOS feature subsets are selected as statistically important subsets for endpoints with CH. Due to clear cytoplasm, CC differs from other subtypes in terms of color, glandular objects and texture. Thus, the C, GT and GOS feature subsets tend to emerge for endpoints with CC. Due to nuclear clusters, PA differs from other subtypes in terms of nuclear properties represented by the NS, NStT, and NStOS feature subsets. Finally, due to its compact nuclear nests, ON differs from CH in terms of topology, represented by the NStTo.

Renal cancer is graded using the Fuhrman nuclear grading system. G1 cells have small intensely stained nuclei with no visible nucleoli. G2 cells have finely granular chromatin, which leads to slightly textured nuclei. They may have inconspicuous nucleoli. In G3, the nucleoli must be easily unequivocally recognizable. G4 is characterized by nuclear pleomorphism (varying size of nuclei), hyperchromasia (leading large nuclear size) and single to multiple macronucleoli [136]. Therefore, nuclear shape

features should be the most important feature set for renal cancer grading. Thus, it is not surprising that the NS, NStOS and NstT subsets are statistically important for most grading endpoints. Literature [136] supports occasional cytoplasmic changes (becomes eosinphilic) in clear cell with higher grade, this is possibly the reason for selection of CStO and CStT subsets for G2 vs. G4 and G3 vs. G4.

Conclusion

In this chapter, we developed a CDSS with a comprehensive set of existing image features that can be applied to a wide variety of histological diagnosis applications. We assessed the predictive performance of the system by applying it to several renal tumor endpoints. We also evaluated the contribution of feature subsets to each disease endpoint in order to reveal emergent properties in the histological images that may relate to biological properties. Results indicate that the feature sets that emerge from the system are biologically interpretable.

CHAPTER 6

BIOLOGICALLY INTERPRETABLE DESCRIPTION OF HISTOPATHOLOGICAL IMAGES

Introduction

Chapter 5 illustrated a comprehensive feature set for histopathological images, evaluated the set on a variety of renal tumor endpoints, and discussed and biologically interpreted emergent feature subsets. As shape-based features were prominent among the emergent subsets, this chapter focuses on the development of novel shape-based features. The proposed shape-based features quantify the distribution of shape patterns in an image using Fourier shape descriptors. The research presented in this chapter was conducted in collaboration with other researchers and most of the content is part of a published article [137].

Over the last decade, several CDSSs have been developed to aid histological cancer diagnosis and to reduce subjectivity. All of these systems attempt to mimic pathologists by extracting features from histological images. Some important features include color, nuclear shape, fractal, textural gray-level co-occurrence matrices (GLCM), wavelets, and topological, among others [5, 138]. Several diagnostic systems for renal cell carcinoma (RCC) are good examples of the utility of these features. For example, Chaudry et al. proposed a system using textural and morphological features with automated region-of-interest selection for RCC subtype classification [36, 139]. Waheed et al. performed a similar analysis but included fractal as well as textural and morphological features [140]. Choi et al. extended the morphological analysis to three-dimensional nuclei and applied

their system to RCC grading [141]. In addition to morphological features, Francois et al. used cell kinetic features in their RCC grading system [142]. Finally, Raza et al. used a scale invariant feature transform (SIFT) method to classify RCC subtypes [143]. Despite the success of these methods in terms of diagnostic accuracy, widespread use of these systems is limited by a lack of feature interpretability. Some researchers have provided visual interpretation of features. For example, some topological features have been related to the amount of differentiation in varying cancer grades [125]. On the other hand, pathologists may not be receptive to, or confident in, features such as wavelet or fractal representations of images because they are not easy to interpret biologically. Moreover, most existing systems exploit morphological properties of nuclear shapes and ignore cytoplasmic and glandular structures despite evidence of their utility [39]. Thus, methods based on a holistic view of shapes and colors may more accurately reflect the process by which a pathologist interprets a renal tumor image [136].

Fourier shape descriptors, described by Kuhl and Giardina [132] have been reported to be very useful as shape descriptors. They are highly robust to high frequency noise because of their ability to reject higher harmonic shape descriptors. Researchers have used Fourier shape descriptors for various medical imaging applications, including shape-based vertebral image retrieval [144], and classification of breast tumors [145]. The medical images involved in these studies typically have definite shapes with consistent landmarks. In addition, researchers have used Fourier shape descriptors for analyzing the shape of a nuclear structure [33, 146, 147]. Histological images, however, lack such landmarks and they tend to exhibit multiple highly variable shapes. As such, it

is difficult to compare histological images using common techniques such as template matching with an image atlas [148] or using shape-based similarity measures after registration of the shapes in a histological image [149]. Therefore, in order to characterize and compare histological images in terms of shapes, we quantify the distribution of shape patterns in an image using Fourier shape descriptors.

We use three steps to build a diagnostic model from a set of histological images: (1) shape-based feature extraction, (2) feature selection, and (3) classifier model selection. We then evaluate this model-building process by examining the biological relevance of shapes (i.e., examining the subtype-specific tissue shapes and cellular structures that correspond to the best features of the classification model) and testing the classifier prediction performance using independent images. Finally, we compare the shape-based diagnostic model to diagnostic models based on traditional histological image features. We show that Fourier shape-based features (1) are capable of classifying H&E-stained renal tumor histological images, (2) out-perform or complement traditional histological image features used in existing automated systems, and (3) are biologically interpretable.

Materials and Methods

Datasets

We perform this study on photomicrographs of H&E stained renal tumor samples. We use two separately acquired datasets: D1 and D2 (Figure 34). D1 contains 48 images

with 12 images of each subtype while D2 has 55 images including 20 CH, 17 CC, 13 PA, and 5 ON subtypes. D2 has samples with nuclear grade varying from 1 to 4.

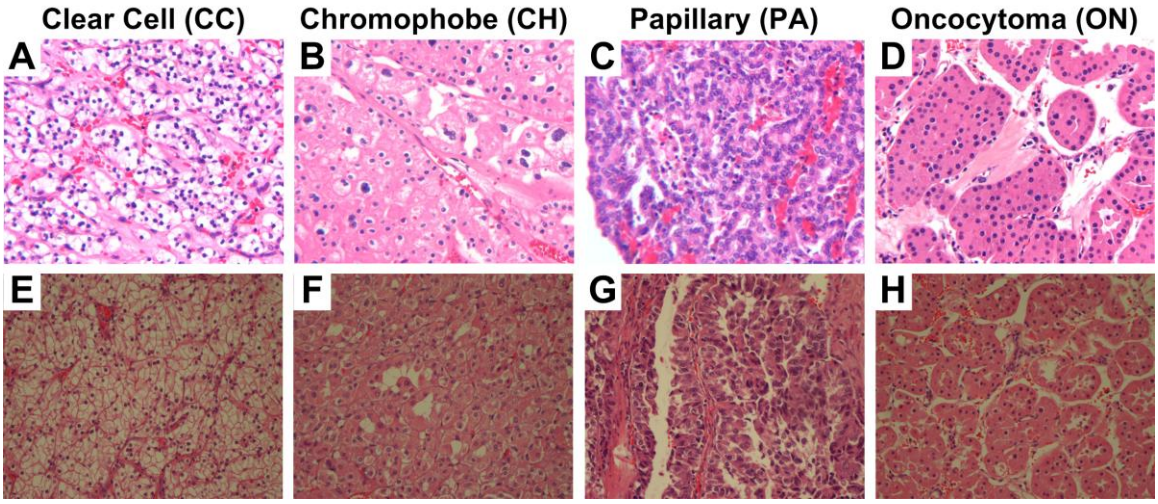

Figure 34: Example images of four H&E stained histological renal tumor subtypes in datasets D1 (A-D) and D2 (E-H).
Among four subtypes, three are renal cell carcinoma (RCC) subtypes: (A and E) clear cell, (B and F) chromophobe, and (C and G) papillary. The fourth subtype is a benign renal (D and H) oncocytoma tumor.

Shape Descriptors

We segment the nuclear, cytoplasmic and no-stain/glandular regions of the original histological image using the automatic color segmentation method described in Chapter 2. Figure 35 illustrates some color segmentation results. *First row*: original histological renal tumor subtype images; *second row*: pseudo colored segmentation masks, where blue, white and pink colors correspond to nuclear, cytoplasmic and no-stain/glandular masks respectively; *third row*: segmented shape contours in nuclear (blue), no-stain/glandular (black) and cytoplasmic (pink) masks.

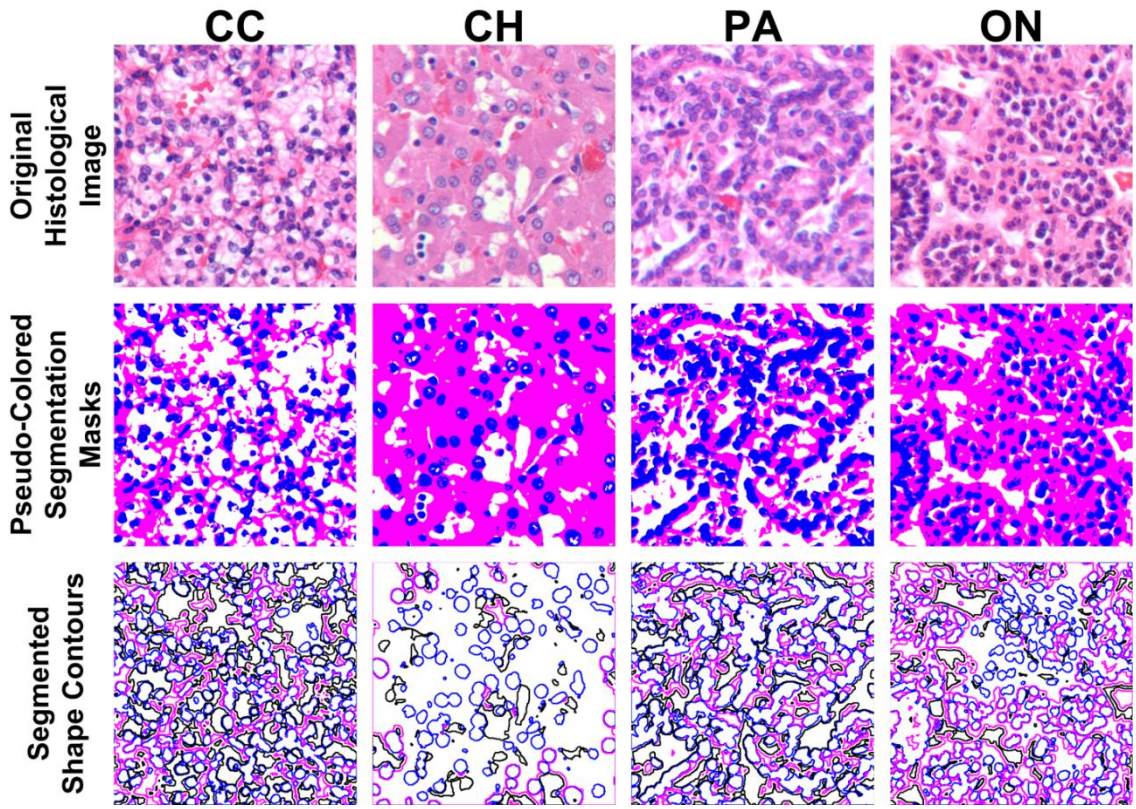

Figure 35: Color segmentation results and shape contours in three masks for four renal tumor subtypes

We represent shape contours using Fourier shape descriptors. Let ($x(t)$, $y(t)$) be parametric representation for each shape contour, then the Fourier series expansion for the one-dimensional periodic function $x(t)$ and $y(t)$ is given by

$$x(t) = A_0 + \sum_{n=1}^{\infty} a_n \cos\frac{2n\pi t}{T} + b_n \sin\frac{2n\pi t}{T} \tag{26}$$

$$y(t) = C_0 + \sum_{n=1}^{\infty} c_n \cos\frac{2n\pi t}{T} + d_n \sin\frac{2n\pi t}{T} \tag{27}$$

where n is the number of harmonics. We estimate the Fourier coefficients A_0, C_0, a_n, b_n , c_n, and d_n by the formulas illustrated in [132]. A_0 and C_0 correspond to the location of a shape, so we do not consider them as shape descriptors. a_n, b_n , c_n, d_n are the shape

descriptors that have commonly been used for shape discrimination [145, 150] and shape retrieval [144, 151] applications in 4*N dimensional space, where N is the number of harmonics. However, we are classifying images based on the distribution of multiple shapes within the images and not based on individual shapes. Therefore, we quantify the distributions of an individual descriptor over all the shapes in an image mask and use these distributions as shape-based features for classification (described in next section). The distribution of four coefficients, a_n, b_n, c_n, d_n for harmonic n, cannot be used separately because they jointly describe an ellipse:

$$x_n(\theta) = a_n \cos\theta + b_n \sin\theta \tag{28}$$

$$y_n(\theta) = c_n \cos\theta + d_n \sin\theta, \tag{29}$$

where $\theta = \frac{2n\pi t}{T}$.

However, using both the semi-major and semi-minor axis lengths of ellipses, we can capture the shape patterns. We quantify semi-major and semi-minor axis lengths as follows. The magnitude of the ellipse phasor is given by

$$r(\theta) = \sqrt{x_n^2 + y_n^2}. \tag{30}$$

We can locate the extrema of this phasor magnitude by differentiating the equation and solving for its root. The resulting solution for θ is

$$\theta_n = \frac{1}{2}\tan^{-1}\left[\frac{2(a_n b_n + c_n d_n)}{a_n^2 + c_n^2 - b_n^2 - d_n^2}\right], \text{where } 0 \le \theta \le \pi \tag{31}$$

Now, as $r(\theta)$ describes an ellipse, θ_n gives the location of either major or minor axis while the other axis is given by $\theta_n + \pi/2$. Therefore, semi-major and semi-minor axes are given by

$$r_n^1 = \max(r(\theta_n), r(\theta_n + \frac{\pi}{2})) \quad (32)$$

$$r_n^2 = \min(r(\theta_n), r(\theta_n + \frac{\pi}{2})) . \quad (33)$$

r_n^1 and r_n^2 capture the magnitude of a shape's variation in the n^{th} harmonic. For n=1, r_n^1 and r_n^2 encode the size of the shape. For n>1, r_n^1 and r_n^2 encodes the complexity of the shape. For simpler shapes, i.e. closer to an ellipse, r_n^1 and r_n^2 quickly reduce to zero with increasing n, while for more complex shapes, they reduce slowly. Figure 36 illustrates shape axes descriptors for synthetically generated clusters of nuclei. In Figure 36.B, for the 1st harmonic, axes features describe size and eccentricity of a shape. For higher harmonics, axis lengths encode detail about the shape. Therefore, in Figure 36.C and Figure 36.D, for 2nd and 3rd harmonics, simple (closer to an ellipse) shapes (such as the green shapes) have axis lengths close to zero, while all other shapes have larger axis lengths.

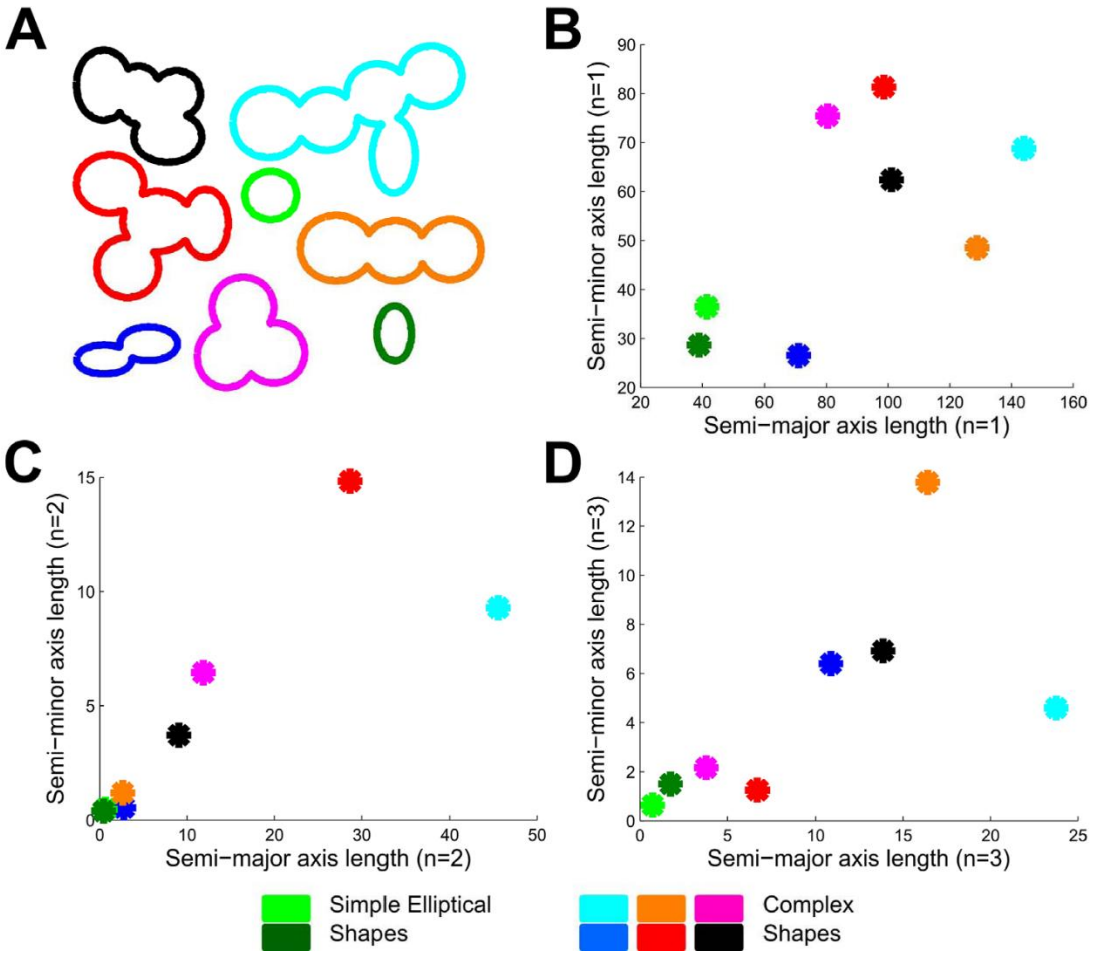

Figure 36: Axis lengths of shape descriptors capture the complexity of shapes
A) We use several synthetic shapes to illustrate the utility of Fourier shape descriptors in capturing shape complexity. The green and light green shapes are the simplest elliptical shapes. B-D) Major and minor axis lengths (in pixels) of the Fourier descriptor ellipses in (A), for harmonics n=1, 2 and 3. Marker colors in (B-D) correspond to shape colors in (A).

Figure 37 illustrates the ability of the axis length distribution to capture the shape profile of an image. In this figure, we are considering nuclear (blue) mask shapes for two RCC subtypes—CH and PA. Figure 37.A and Figure 37.D are histograms of major axis length at harmonic two. The y-axis of the histogram is the frequency of shapes with a particular range of coefficient value. The second harmonic captures the complexity of the shape approximation. Thus, for complex shapes like PA's nuclear clusters, the major axis length of the second harmonic tends to have higher values compared to that of simpler shapes like individual circular nuclei. In Figure 37.C and Figure 37.F—corresponding to the histograms in Figure 37.A and Figure 37.D, respectively—we have outlined, in cyan, shapes with values of major axis length that fall the in lower seven bins. Shapes with values of major axis length falling in the upper eight bins are outlined in blue. We can observe that the CH image (Figure 37.A-C) has a dominant pattern of simple shapes as compared to the PA image (Figure 37.D-F). As described in the next section, discretization of axis lengths of all the shapes in an image is the basis for representing a histopathological image as a multi-feature observation.

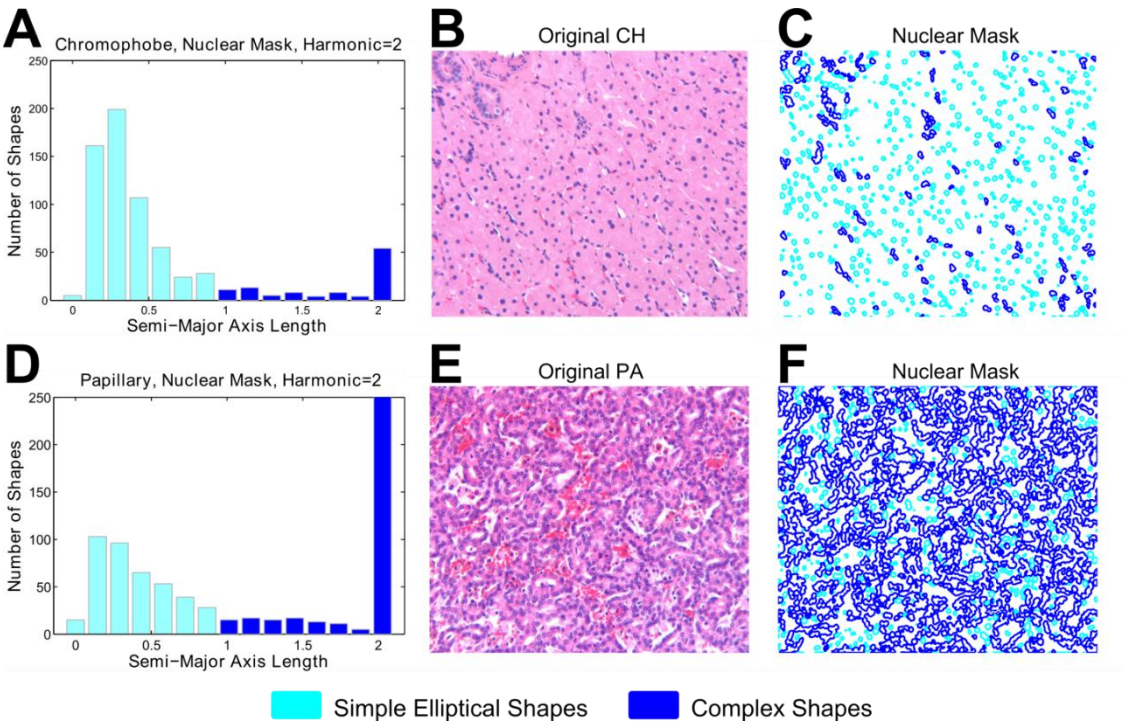

Figure 37: Fourier shape features discriminate simple and complex shapes in histological images.

The bar graphs illustrate the distribution of the second harmonic's major axis length of all the shapes in the nuclear mask for (A) a chromophobe and (D) a papillary image. (B-C) and (E-F) are original image and nuclear mask shapes of chromophobe and papillary, respectively.

Discretization of Shape Descriptors

In order to develop a classification system, we represent each image as a single observation with a fixed number of features. Due to the variable number of shapes in each image, we quantify the distribution of shape descriptors (axis lengths) to create a "shape profile", represented as a histogram. We determine the dynamic range of each histogram by computing interquartile distances of shape descriptor distributions from the training set. Interquartile distance is the distance between the 25th and 75th percentiles of a distribution [152]. Mathematically, $R_n^{c,m}$ is the distribution of axis lengths over all shapes in all images in the training dataset for a particular combination of harmonic (n), axis type (c) and mask (m). Let function $f_P(R)$ return the p^{th} percentile of distribution R, then the interquartile distance (IQD) is given by

$$IQD(R) = f_{0.75}(R) - f_{0.25}(R) \tag{34}$$

Using equation (5), we $R_n^{c,m}$:

$$L_n^{c,m} = \max(0, (f_{0.5}(R_n^{c,m}) - 2 * IQD(R_n^{c,m}))) \tag{35}$$

$$U_n^{c,m} = f_{0.5}(R_n^{c,m}) + 2 * IQD(R_n^{c,m}) \tag{36}$$

where L, U are the lower and upper bounds of the range, respectively. Outliers bin into the edges of the histogram and may be informative features. Axis lengths are always positive, therefore the lower bound of the range is forced to be greater than or equal to zero. Figure 38 illustrates the data flow from a histological RGB image to a list of 900 features. The procedure is as follows:

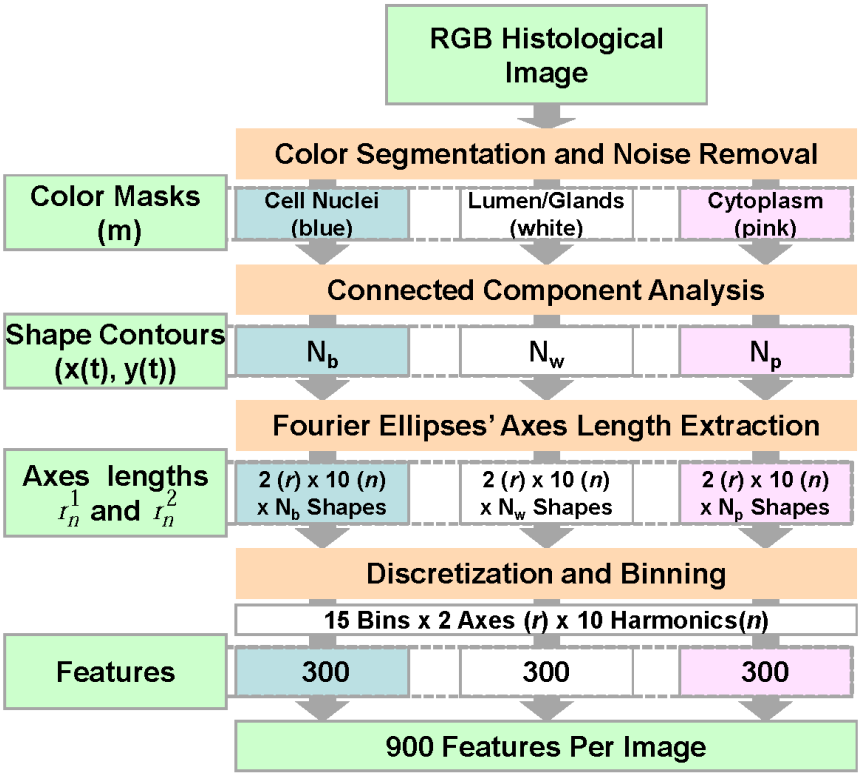

Figure 38: The data flow for extraction of 900 shape-based features from a RGB histological image.

For each mask, we obtain the contour for all shapes after noise filtering using connected component analysis. N_m is number of shapes in m mask, where $m \in \{b, w, p\}$. We then extract shape axes descriptors (2 axes*10 harmonics) for each shape contour and bin them to produce 2*10 histograms for each mask (3 masks*10*2 histograms in an image). Due to the variation in dynamic range of the two axes and harmonics, we use data-dependent histogram ranges with 15 bins per histogram. We use the histogram frequencies as features for our image classification.

1. Generate a binary mask for each color in the histological image. We use three colors for H&E stained RCC images: blue (nuclear), white (no-stain/glandular), and pink (cytoplasmic).

2. Extract contours for all shapes in a mask after connected component analysis.

3. Extract axis lengths for Fourier ellipses (r_n^1 and r_n^2) for the first 10 harmonics (n). This will give us 2*10 variables for each shape.

4. For each harmonic (n), axis type (c), and mask (m), perform a binning procedure (Figure 38). We generate 20 histograms for each mask. We use 15 bins and a range determined by $L_n^{c,m}$ and $U_n^{c,m}$ as previously described.

5. Combine histogram frequency from the three masks to generate a list of 900 shape-based features

There are a number of advantages in using discretization rather than Euclidian distance to compare images. First, the axes of shapes that are similar, but perhaps not identical, fall into the same histogram bin. Similar histogram frequencies can be interpreted as a similarity of shapes between images. Second, bins sensitive to noise or outlier shapes in any sample will be rejected during feature selection. Finally, discriminating features can be components corresponding to multiple types of shapes rather than components corresponding to the most prominent characteristic shape.

Traditional Features

Traditional features in computer-aided diagnosis include texture, morphological, topological, and nuclear. In order to compare shape-based features to these traditional features, we extract additional features from histological renal tumor images.

For texture, we have two sets of features: GLCM and wavelet. For GLCM features, we extract a 16 × 16 GLCM matrix for each gray-scale tissue image with 16 quantization levels [127]. Using this matrix, we extract 13 texture properties including contrast, correlation, energy (angular second moment), entropy, homogeneity (inverse difference moment), variance, sum average, sum variance, sum entropy, difference variance, difference entropy, and two information measures for correlation. These features are reported to successfully capture texture properties of the image and are very useful in automated cancer grading [125, 127, 153].

For wavelet features, we perform three-level wavelet (db6) packet decomposition [130] of the gray-level tissue image and extract energy and entropy [38] of 84 coefficient matrices (level 1, 2 and 3), producing 168 features. Wavelet features capture texture properties of an image.

For morphological features, we use color-GLCM, a method proposed by Chaudry et al. to classify renal tumor subtypes. This method generates a four-level gray-scale image from four color stains in H&E-stained images [36]. The four colors resulting from H&E-stained images (blue, white, pink, and red) correspond to segmented regions of nuclei, lumen, cytoplasm, and red blood cells. We then extract a 4 × 4 GLCM matrix for

the gray-scale image. We extract 21 features from this matrix including 16 elements of the 4 × 4 GLCM matrix, contrast, correlation, energy (angular second moment), entropy, and homogeneity (inverse difference moment). These features capture morphological features of the image such as stain area and stain co-occurrence properties.

For topological features, we use a graph-based method. Several researchers have proposed graph-based features to capture the distribution of patterns in an image. Biologically, these features capture the amount of differentiation (related to cancer grade) in a histological image. We morphologically erode our nuclear mask to separate nuclear clusters and use their centroids (nuclear centers) for this analysis. First, we create a Voronoi diagram from these centers and then calculate area and perimeter of each region and all side-lengths. We then calculate mean, minimum, maximum, and disorder of the distribution to produce 12 features [125]. The disorder, D, of a distribution, r, is given by

$$\mathrm{D}(r) = 1 - \left(1 + \sigma_r / \mu_r\right)^{-1}, \tag{37}$$

where σ_r and μ_r are standard deviation and mean of r, respectively [56]. Second, we calculate the area and side lengths of the Delaunay triangles and extract statistics similar to those of the Voronoi diagram to produce eight more features. Last, we calculate side lengths of the minimum spanning tree and extract the same statistics to produce four more features. In total, we extract 24 topological features.

For nuclear features, we extract nuclear count and elliptical-shape properties, which have proven to be useful for renal carcinoma subtyping and grading. For segmenting nuclear clusters, we use an edge-based method with three steps: concavity detection,

straight-line segmentation, and ellipse fitting. We describe each elliptical nucleus using area, major-axis length, minor-axis length, and eccentricity. We then calculate mean, minimum, maximum and disorder of the distribution of these descriptors to produce 16 features. In total, including nuclear count, we extract 17 nuclear features.

We combine the GLCM (13 features), color-GLCM (21), wavelet (168), topological (24), and nuclear (17) features to produce a set of 243 "Combined Traditional" features. Finally, we combine the "Combined Traditional" (243) and "Shape" (900) features to a produce a set of 1143 "All" features.

Feature Selection and Classification

For validation, we combine D1 and D2, then randomly split them into two new training and testing datasets with balanced sampling from both datasets. We perform a three-fold split, in which two folds form the training set while one fold forms the testing set. Each fold acts as a testing set once, resulting in three training–testing sets. We perform 10 iterations of this split to estimate the variance in performance. Thus, there are 30 training–testing sets in the external CV that produces the final classification accuracy. For each of the 30 training sets, we perform an additional three-fold, 10 iterations of CV to choose an optimal set of classifier and feature selection parameters. This forms the internal CV of a nested CV (Appendix).

We construct a multi-class classification system consisting of a hierarchy of binary classifiers CC vs. PA, CC vs. CH, CC vs. ON, CH vs. PA, CH vs. ON, and ON vs. PA

also called a directed acyclic graph (DAG) classifier [154]. According to Platt et al., the order of binary comparisons has little effect on the overall classification accuracy. Thus, we use the hierarchy illustrated in Figure 39.

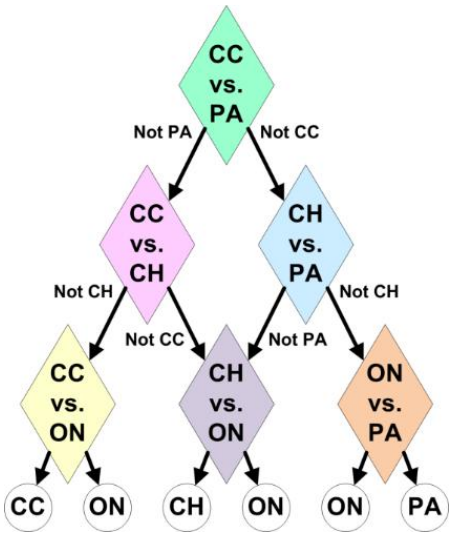

Figure 39: A multi-class hierarchy of binary renal tumor subtype classifiers, also known as a directed acyclic graph (DAG) classifier.

Each node in the hierarchy is independently optimized such that, for each binary comparison, we choose a set of model parameters (i.e., classifier as well as feature selection parameters). We consider 224 SVM classifier models including 14 kernel types (linear or radial with the gamma parameter ranging from 2^{2}, 2^{1}, $2^{0}$to 2^{-10}) and 16 cost values (2^{-5}, 2^{-4}, 2^{-3} to 2^{10}) [155, 156]. We considered the following feature sizes for different features (e.g., starting feature size : feature step size : ending feature size):

1. GLCM (1:1:13)
2. Color-GLCM (1:1:21)
3. Wavelet (1:5:166)

4. Topological (1:1:24)
5. Nuclear (1:1:17)
6. Combined Traditional (1:6:243)
7. Shape and All (5:5:180)

We choose the feature size step such that the total number of feature sizes is approximately 40. For Shape and All features we also consider number of harmonics (n=2 to 10) as a feature selection parameter. We choose the simplest model with a CV accuracy within one standard deviation of the best performing model [157]. In choosing the simplest model, we give preference to the linear SVM kernel over the radial SVM kernel and lower values of gamma for the radial SVM kernel, SVM cost, number of harmonics, and feature size.

We select features using mRMR, which selects a set of features that maximizes mutual information between class labels and each feature in the set; and minimizes mutual information between all pairs of features in the set) [93]. Our features are continuous and, as suggested by Ding et al., we use Mutual Information Quotient (mRMR-q) optimization after discretization using the following transform:

$$k' = \begin{cases} -1 & k < \mu_k - \sigma_k / 2 \\ 0 & \mu_k - \sigma_k / 2 \le k \le \mu_k + \sigma_k / 2 \\ 1 & k > \mu_k + \sigma_k / 2 \end{cases} \quad (38)$$

where k' is the transformed feature k, μ_k and σ_k are the mean and standard deviation of feature k over all samples in the training dataset, respectively.

Results and Discussion

Prediction Performance in Renal Subtyping

Fourier shape-based features are capable of classifying histological renal tumor subtype images with high accuracy and simple classification models. Table 10 lists the shape-based prediction performance of the multi-class renal tumor classifier (using a Directed Acyclic Graph, DAG, classifier) as well as that of each binary comparison (discrimination of every pair of subtypes). The shape-based multi-class classifier predicts the subtypes of renal tumor images with an average accuracy of 77%. The average prediction accuracy for each binary comparison ranges between 83%-96%.

Table 10: Predictive performance of Fourier shape-based features.

Endpoint	Inner CV Accuracy	External CV Accuracy
DAG	N/A	0.77±0.03
CH vs. CC	0.83±0.03	0.83±0.05
CH vs. ON	0.83±0.02	0.84±0.04
CH vs. PA	0.97±0.01	0.96±0.02
CC vs. ON	0.90±0.02	0.90±0.07
CC vs. PA	0.96±0.01	0.95±0.04
ON vs. PA	0.94±0.01	0.93±0.04

Figure 40 shows that the optimal classifier parameters correspond to fairly simple models. The parameters are optimized by selecting the simplest model with predictive performance within one standard deviation of the highest performing model. We define a simple model as one that has small feature size, low SVM cost, low SVM gamma (or is a linear SVM), and that prefers features from smaller Fourier shape descriptor harmonics.

Although we consider ten harmonics of shape-based features, parameter selection usually only selects the first few (< 4) harmonics (Figure 40.A). In Figure 40.A, even though the best choice for number of harmonics varies for different binary classifiers, all binary classifiers have a decreasing trend of selecting a high number of harmonics. Figure 40.B illustrates the distribution of feature sizes selected for binary classifier models. We can observe that there is a decreasing trend of selecting large feature sizes and, in most cases, less than 20 features were selected. Figure 40.C and Figure 40.D show that optimal SVM cost and SVM gamma selections are also low, with preference given to linear SVMs over non-linear radial basis SVMs. Table 11 lists the most frequently selected parameters for each of the six binary classifiers.

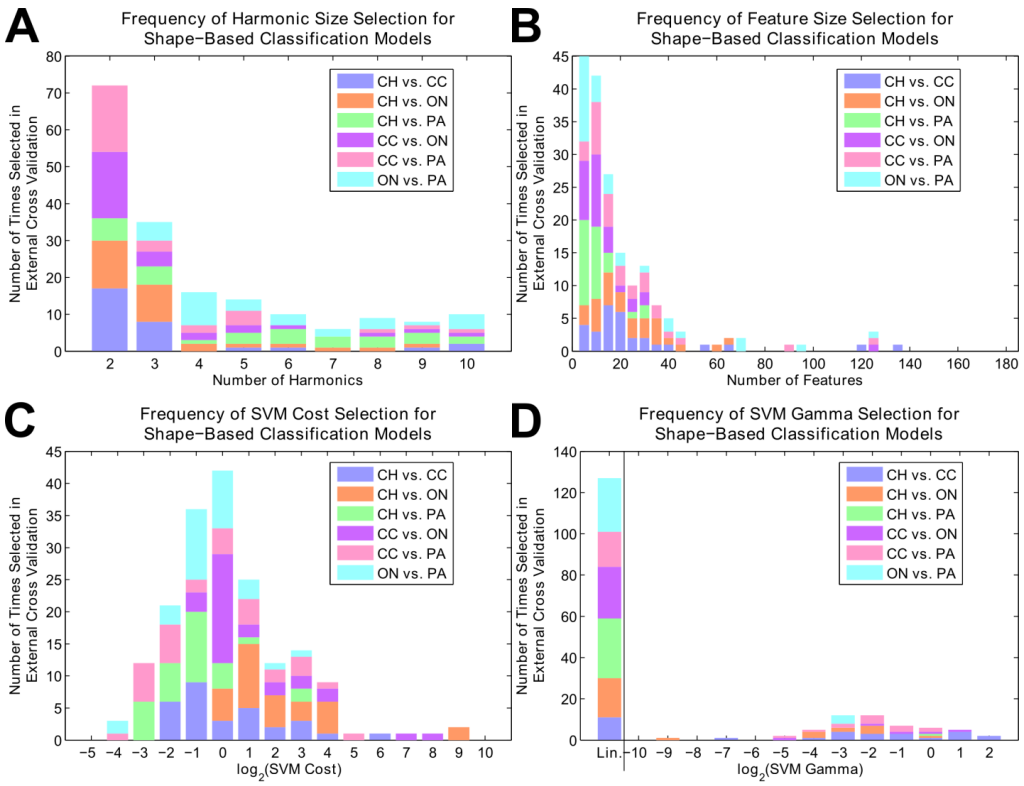

Figure 40. Parameter space investigation for shape-based classification models. Distribution of a) maximum number of harmonics considered, b) number of features selected, c) SVM cost selected, and d) SVM gamma selected over 10 iterations times three folds of parameter estimation for various binary endpoints. The total length of each bar is a summation of counts over all binary endpoints while each color indicates count for a specific binary comparison.

Table 11: Frequently selected parameters for binary, shape-based renal-tumor subtyping models.

Binary Endpoint	Harmonic	Feature Size	SVM Cost*	SVM Gamma
CH vs. CC	2	15	-1	Linear
CH vs. ON	2	10	1	Linear
CH vs. PA	2	5	-1	Linear
CC vs. ON	2	10	0	Linear
CC vs. PA	2	10	-3	Linear
ON vs. PA	4	5	-1	Linear

* Log_2values

We use nested CV to select prediction model parameters and to evaluate these prediction models on independent data. The nested CV procedure includes 10 iterations of three-fold external CV with 10 iterations of three-fold internal CV. Although there is some variance across the iterations of CV, Figure 41 shows that mean internal CV is a good estimate of mean external CV for each of the binary comparisons. Each point in Figure 41 corresponds to an iteration of external CV for each binary comparison. The horizontal position of each point is internal CV accuracy averaged over 10 iterations and three folds. The vertical position of each point is external CV accuracy averaged over three folds. Classifier model parameters for each point are selected from among 72,576 models consisting of 36 feature sizes, 14 types of SVM classifiers (linear SVM and radial basis SVM classifiers over 13 different gammas), 16 SVM cost values, and 9 values for the number of harmonics. The optimal parameter set for each classifier model corresponds to the simplest model (i.e., smallest feature size, smallest cost, smallest gamma, and smallest number of harmonics) within one standard deviation of the best performing model. This high concordance of internal CV and external CV performance indicates that internal CV performance is predictive of external CV performance and

classifier models generated from shape features are robust and will perform similarly for future samples. Moreover, the binary comparisons discriminating CH vs. PA, CC vs. ON, CC vs. PA, and ON vs. PA tend to result in high performance (> 90%) while the binary comparisons discriminating CH vs. CC and CH vs. ON result in moderate performance (~83-84%). We describe the reasons for these observations below.

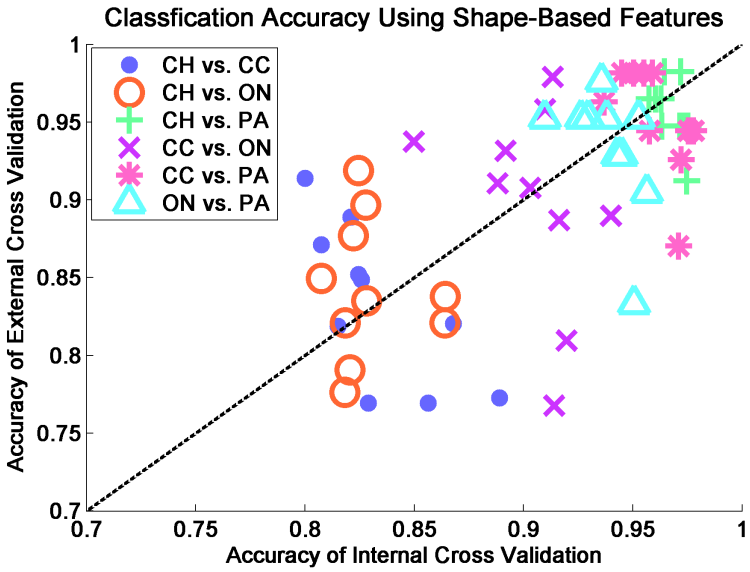

Figure 41: Scatter plot of inner CV vs. external CV average validation accuracy values, during 10 external CV iterations, for six pair-wise renal tumor subtype comparisons.

CC and PA are the most prevalent subtypes of RCC and are generally the easiest for pathologists to visually identify. Consequently, discriminating shape-based features for these classes are easy to identify, resulting in high classification performance. One exception, however, is the CH vs. CC comparison. CH is known to exhibit some CC

properties such as clear cytoplasm. As a result, the prominent feature for the CC subtype is sometimes not sufficient for accurate classification of CC and CH. Moreover, the ON renal tumor subtype is histologically and genetically very similar to the CH RCC subtype, despite the fact that ON is a benign tumor whereas CH is a carcinoma [158]. This similarity explains the moderate performance of the CH and ON binary classifier.

Comparison with Traditional Histopathological Features

Table 12 shows that, in comparison to five traditional feature sets, classification of renal tumor subtypes based on shape-based features performs well. In fact, the performance of shape features is similar to the combined traditional features, which includes texture, topological and nuclear properties. In some cases, combining shape-based features with traditional features (i.e., 'All' features) improves prediction performance, indicating that shape-based features can complement traditional features. Table 12 lists the means and standard deviations over 10 iterations of external CV for each binary comparison as well as for the multi-class DAG classifier.

Table 12: External CV accuracy of renal binary subtyping models using Fourier shape vs. traditional features.

Endpoint	GLCM	Color GLCM	Wavelet	Topological	Nuclear	Combined Traditional	Shape (Proposed)	All
DAG	0.57±0.04	0.67±0.02	0.52±0.06	0.50±0.03	0.66±0.03	0.79±0.04[a]	0.77±0.03[b]	0.78±0.03[c]
CH vs. CC	0.75±0.05	0.77±0.05	0.74±0.05	0.74±0.05	0.76±0.06	0.81±0.03	0.83±0.05	0.82±0.05
CH vs. ON	0.76±0.05	0.68±0.06	0.67±0.05	0.72±0.05	0.79±0.05	0.86±0.05	0.84±0.04	0.88±0.04
CH vs. PA	0.85±0.04	0.95±0.02	0.86±0.05	0.80±0.04	0.91±0.04	0.94±0.02	0.96±0.02	0.96±0.03
CC vs. ON	0.74±0.06	0.78±0.06	0.63±0.03	0.77±0.07	0.93±0.04	0.93±0.04	0.90±0.07	0.91±0.05
CC vs. PA	0.78±0.06	0.97±0.04	0.69±0.09	0.59±0.07	0.76±0.07	0.95±0.05	0.95±0.04	0.97±0.03
ON vs. PA	0.74±0.07	0.86±0.06	0.74±0.07	0.65±0.04	0.96±0.03	0.97±0.03	0.93±0.04	0.92±0.04

Note: The difference between a, b, and c is not statistically significant; p-values of the null hypothesis between a & b; a & c; and b & c using a t-test are 0.16, 0.70, and 0.23, respectively.

Figure 42 shows the contribution of each feature type to the classification model when considering 'All' features. The box plots in Figure 42 represent the distribution of percent contribution of each feature type to a binary classifier over 10 iterations of external CV. We can make the following observations from Figure 42: 1) Shape features have a high (>55%) contribution for all binary endpoints, which indicates that the feature selection method ranks shape features higher than other features. The contribution is comparatively lower for CH vs. CC, CH vs. ON, and CC vs. ON endpoints because other traditional features were also useful for these endpoints. 2) Nuclear features, which capture nuclear-shape properties, highly contribute to all six endpoints 2) In addition to shape features for the CH vs. ON endpoint, topological, nuclear and wavelet features also contribute to the prediction models, resulting in a 4% increase in accuracy compared to shape features alone. This indicates that, in addition to shape (Fourier and nuclear) properties, CH and ON differ in topological and wavelet properties. 3) Color GLCM

performs very well for CC vs. PA classification. Thus, color GLCM is a major contributor for CC vs. PA classification, resulting in a 2% increase in accuracy.

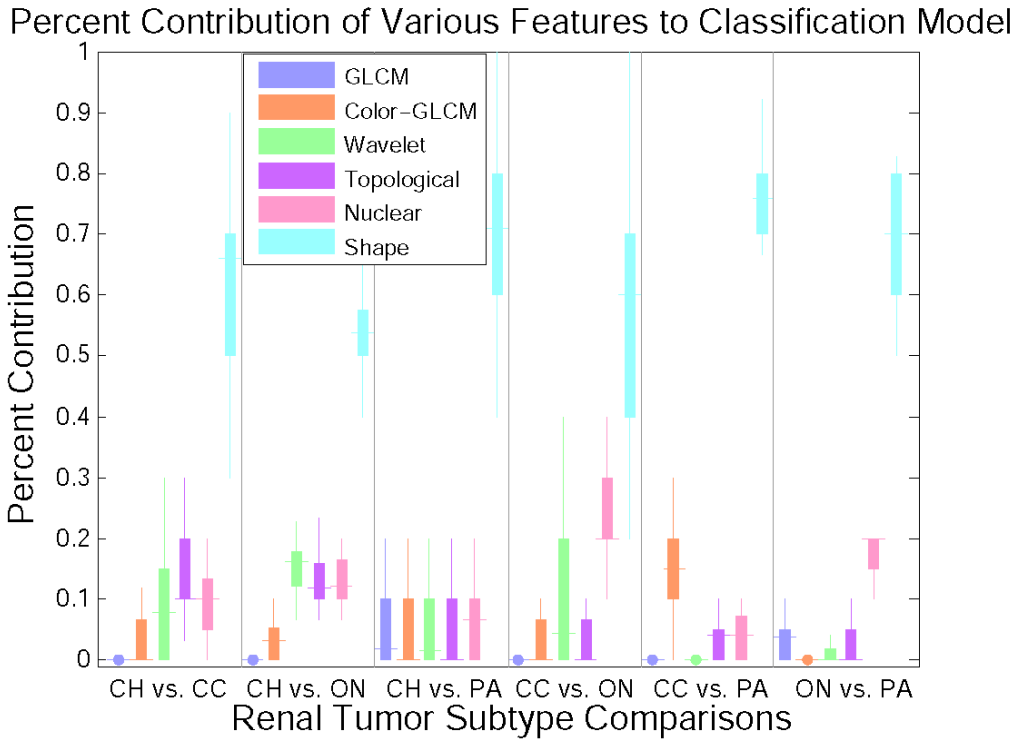

Figure 42: Renal tumor binary classification models use a variety of features to quantify important biological properties.
Percentage contribution of different features for each binary comparison in 'All' features model. The contribution of shape features tends to be greater than 55% for all endpoints (median value, marked by horizontal line).

Biologically Interpretation

Figure 43 illustrates the biological interpretability of shape-based features for each renal tumor subtype. In order to visualize the biological significance of the features identified by our feature selection method, we overlay the top discriminating shapes on the images of renal tumor subtypes for each binary comparison. Feature selection identifies individual shape axes and not entire shapes. Thus, discriminating shapes are shapes with axes values that have been discretized into a bin corresponding to a highly ranked feature.

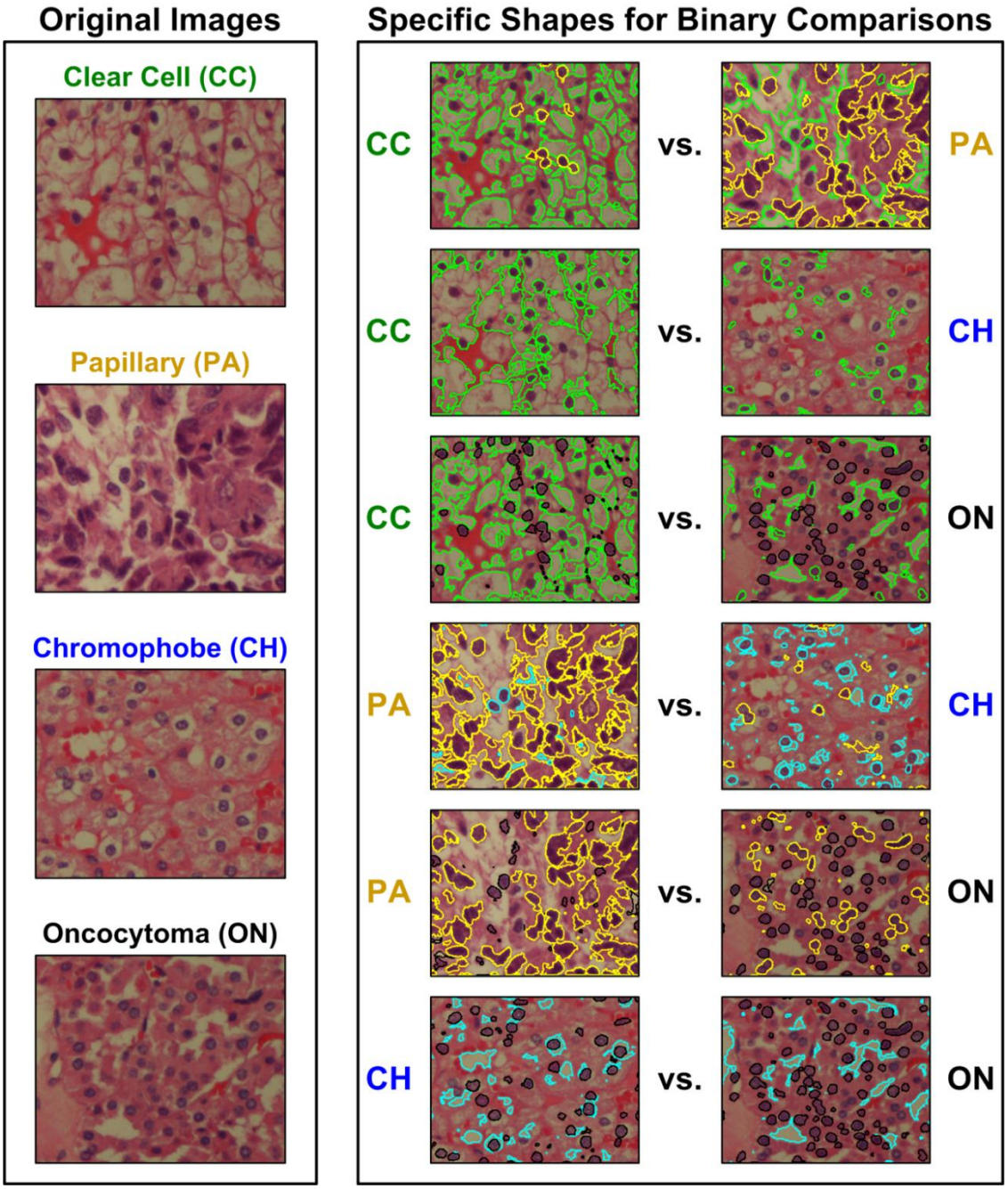

Figure 43: The top discriminating shapes for six binary endpoints correspond to pathologically significant shapes in histological renal tumor images.
Green shapes: occur more frequently in clear cell; *yellow shapes*: occur more frequently in papillary; *blue shapes*: occur more frequently in chromophobe; and *black shapes*: occur more frequently in oncocytoma.

For each binary comparison, we identify all shapes in an image that have Fourier axes values corresponding to the top 25 features. We selectively color the shapes based on "over expression", or increased relative frequency for particular subtypes. Shapes highlighted in green occur more frequently in clear cell; yellow shapes occur more frequently in papillary; blue shapes occur more frequently in chromophobe; and black shapes occur more frequently in oncocytoma. Here, we interpret the biological significance of highlighted shapes for each binary comparison.

Histopathological features of the clear cell subtype include clear cytoplasm, compact alveolar, tubular, and cystic architecture leading to distinct cell membranes. Comparing clear cell to papillary and oncocytoma, we see that clear cytoplasm (no-stain/glandular (white) mask region, outlined with green) is the primary distinguishing characteristic that is noticeably less frequent in papillary and oncocytoma. On the other hand, because chromophobe images tend to also exhibit halos resembling clear cytoplasm, the distinguishing features between CC and CH are distinct cell membranes (small cytoplasmic (pink) mask areas outlined with green between larger no-stain/glandular (white) mask areas) that are more frequent in CC compared to CH. Similarity in halos and clear cytoplasm shapes is possibly the reason for low accuracy in the CH vs. CC binary classification. Features of the papillary subtype include scanty eosinphilic cytoplasm and a papillary (i.e., finger-like) pattern of growth resulting in long, complex clusters of nuclei. In all comparisons with the papillary subtype, complex clusters of nuclei are the dominant distinguishing feature and are generally more prominent in papillary (nuclear (blue) mask areas outlined with yellow). The frequency

of nuclear shapes in oncocytoma appears to be similar to that of papillary. However, the nuclear clusters in papillary are generally larger and more irregular due to the clustering, resulting in different Fourier shape axes values. Histopathological features of the chromophobe subtype include wrinkled nuclei with perinuclear halos. When comparing chromophobe to papillary or oncocytoma, our feature extraction and selection method identifies these halos (no-stain/glandular (white) mask areas, outlined with blue). In addition, single nuclei become dominant when comparing with papillary.

Histopathological features of the chromophobe subtype include granular cytoplasm with round nuclei, usually arranged in compact nests or microcysts. These round nuclei appear to be dominant in ON, when comparing to other subtypes. It can be observed that dominant features for both CH and ON are present in the opposite subtype as well. Hence, the difficulty in distinguishing the two subtypes.

Limitations and Computational Complexity

Some limitations of shape-based features for histological image classification depend on the specific biological application. Shape-based features may not be suitable for cases in which the primary discriminating features are not based on shapes. For example, in cancer grading applications, topological and texture properties may be more useful than shape-based features. Moreover, as we have seen the results of Table 12 and Figure 43, shape-based features may not capture all of the important distinguishing information. For example, in the case of the CH vs. ON endpoint, the addition of texture and wavelet features to shape-based features increases prediction performance by 4%. In

addition, for the CC vs. PA endpoint, inclusion of the GLCM texture features increases prediction performance by 2%. Thus, shape-based features are limited to clinical prediction applications that are inherently shape-based, but, in such cases, may be used to complement other non-shape-based features.

The computational complexity of shape-based features is higher than those of traditional histological feature extraction and analysis methods, but should not prevent implementation in a clinical setting. To convert a RGB histological image (1600x1200 pixel portions) into 900 shape-based features (Figure 38), a desktop computer (Intel Xeon E5405 quad-core processor, 20 GB RAM) requires an average of 74.96 seconds. Compared to some histological image features, this processing time is high. However, the processing time depends on the number of harmonics used for representation and the number of shapes in an image. We have reported the processing time for extracting features from the first ten harmonics. However, in practice, we have observed that all optimized models use less than five harmonics. Optimization of these parameters to identify a predictive model can be time consuming depending on the size of the training set. However, in a clinical setting, such a model would only need to be optimized once, and then periodically updated with new patient data. In a clinical scenario, a pathologist that requires a histological diagnosis for a patient would submit a few image samples from a tissue biopsy to a pre-optimized prediction system. Computational time for processing and predicting based on these image samples would be negligible compared to time required for biopsy, image acquisition, and consultation with a pathologist.

Conclusion

In this chapter, we presented a novel methodology for automatic clinical prediction of renal tumor subtypes using shape-based features. These shape-based features describe the distribution of shapes extracted from three dominant H&E stain colors in renal tumor histopathological images. We evaluated the four-class prediction performance of shape-based classification models using 10 iterations of three-fold nested CV. The overall classification accuracy of 77% (average external CV accuracy) is favorable compared to previous methods that use traditional textural, morphological, and wavelet-based features. Moreover, results indicate that combining shape-based features with traditional histological image features can improve prediction performance. The biological significance of the characteristic shapes identified by our algorithm suggests that this automatic diagnostic system mimics the diagnostic criteria of pathologists. We applied this methodology to renal tumor subtype prediction. However, the methodology may be extended to any histological image classification problem that traditionally depends on visual shape analysis by a pathologist. Moreover, these shape-based features may be coupled with other image features to achieve higher diagnostic accuracy.

CHAPTER 7

NORMALIZATION METHODS FOR BATCH-INVARIANT DECISION MAKING

Introduction

This chapter provides validation of information extraction methods on sub-sections of WSIs and addresses the informatics challenge of developing batch-invariant informatics methods. It focuses on information-level batch-effects that affect the prediction performance of CDSSs based on images from multiple institutions or set-ups. The research presented in this chapter was conducted in collaboration with other researchers and most of the content is part of research articles on batch-invariant decision making [159, 160].

CDSSs can guide pathologists in diagnosing cancer by extracting and modeling quantitative properties of histopathological images [5]. However, when histopathological images are acquired in different experimental setups and tested on pre-trained diagnostic models, the prediction performance can suffer due to batch effects, i.e., non-biological experimental variations such as age of sample, method of slide preparation, specifications of the microscope, and type of post-processing software [161]. Batch effects may lead to large differences in quantitative image features. Thus, it is difficult to accurately diagnose patients using prediction models trained with a separate batch. Because of batch effects, a pathology lab that uses multiple imaging devices (e.g., microscopes with mounted digital cameras or whole-slide scanners) may need to maintain multiple diagnostic models.

Moreover, data acquired using older devices or experimental setups cannot be used in training models for future data acquired with newer devices/setups. This poses a huge challenge for cross-laboratory adoption and standardization of CDSSs for pathology.

Batch effects are also a major challenge for other biomedical data modalities. Although the causes of batch effects are different for each data modality, methods developed to remove batch effects may be applicable to multiple data modalities. For example, the sources of batch effects in microarray gene expression data include platform, laboratory, sample preparation protocol and reagents, technician and atmospheric ozone level [26, 27]. Batch effects generally affect the mean (location) and variance (scale/spread) of the data [162]. Therefore, batch effect removal methods focus on normalization of location and scale, e.g., ratio-based methods and ComBat [27, 162]. Luo et al. compared several batch effect removal methods for microarray data and found that ratio-based methods performed the best [27]. In a separate study, Chen et al. compared six batch effect removal methods and found that ComBat performs the best [26].

Removal of batch effects in histopathological images is a relatively new area of research [161]. However, with the emergence of large image data repositories such as TCGA, batch effects have become an increasingly important area of research. Histopathological image analysis studies have primarily focused on single-batch data [37-39, 55]. Some studies have highlighted color batch effects in histopathological images and suggested color normalization methods [21, 22]. Color batch effects, which lead to

variation in stain colors across batches, affect the performance of color segmentation methods and color features. Kothari et al. studied scale batch effects in histopathological images and suggested a scale normalization method based on nuclear area. To the best of our knowledge, no published work quantifies histopathological image batch effects or compares batch effect removal methods for histopathological images.

We compare six normalization methods including one image (scale) normalization method and five feature normalization methods: mean, rank, ratio, ComBatP, and ComBatN. Using four renal tumor histopathological datasets acquired using different experimental setups, we assess the impact of each batch effect removal method on image-based features and downstream prediction of renal tumor subtype and grade. Results indicate that data batch can be a larger source of variance in image features compared to biological factors such as grade and subtype. Most batch effect removal methods can reduce this variance to nearly zero. Moreover, batch effect removal methods can increase cross-batch and combined-batch prediction performance, with ComBatN performing the best.

Materials and Methods

Data

We use digital micrographs of renal tumor biopsy samples acquired in four experimental setups. Tissues samples in all four batches RCC1, RCC2, RCC3 and RCC4 were biopsied and fixed at Emory University. The micrographs for the first three batches were acquired at Emory University while the fourth batch was acquired at the Georgia

Institute of Technology. Table 13 lists image acquisition details for four batches. Each image is a rectangular section manually selected from a WSI by a pathologist. Batches RCC1, RCC3, and RCC4 are annotated with both grade and subtype while batch RCC2 is annotated with only subtype.

Table 13: Image-acquisition devices and parameters for four renal batches.

Factors	**RCC1**	**RCC2**	**RCC3**	**RCC4**
Microscope	Nikon Eclipse 80i	Nikon Eclipse 80i	Olympus BX51	Zeiss Axio Imager z2
Magnification	20x	20x	40x	40x
Camera	Nikon DS-2MV	Nikon DS-2MV	Olympus DP71	Zeiss AxioCam MRm
CCD pixel size (µm)	4.40 x 4.40	4.40 x 4.40	4.40 x 4.40	6.45 x 6.45
Year	2008	2006	2011	2012
Subjects	15	12	12	18
Images	53	36	72	160
Image Format	JPEG	JPEG	PNG	PNG
Image size (Pixels)	1600x1200	1600x1200	2040x1536	2040x1536

Images in these datasets represent one of three prominent renal tumor subtypes—CH, CC, and PA—and Fuhrman grade of one to four. Table 14 lists the number of samples per dataset for each subtype and grade.

Table 14: Distribution of subtypes and grades in four renal carcinoma batches.

		RCC1	**RCC2**	**RCC3**	**RCC4**
Subtype	CH	20	12	26	30
	CC	17	12	24	82
	PA	16	12	22	48
Grade	G1	13	N/A	18	20
	G2	13	N/A	18	63
	G3	13	N/A	18	50
	G4	14	N/A	18	27

Figure 44 illustrates 512x512-pixel subsections of three subtype samples in each of the four batches. These subtypes are histopathological subtypes and can be visually predicted based on morphology [136]. CC has clear cytoplasm with distinct cell membranes and round nuclei. CH has granular cytoplasm with prominent cell membranes, wrinkled nuclei, and perinuclear halos (i.e., white stain surrounding nuclei). PA has finger-like complex nuclear clusters. These properties are visually apparent in each of the four batches. However, images of each subtype appear very different between batches. In particular, RCC1 and RCC2 images appear to have different texture and scale compared to RCC3 and RCC4 images.

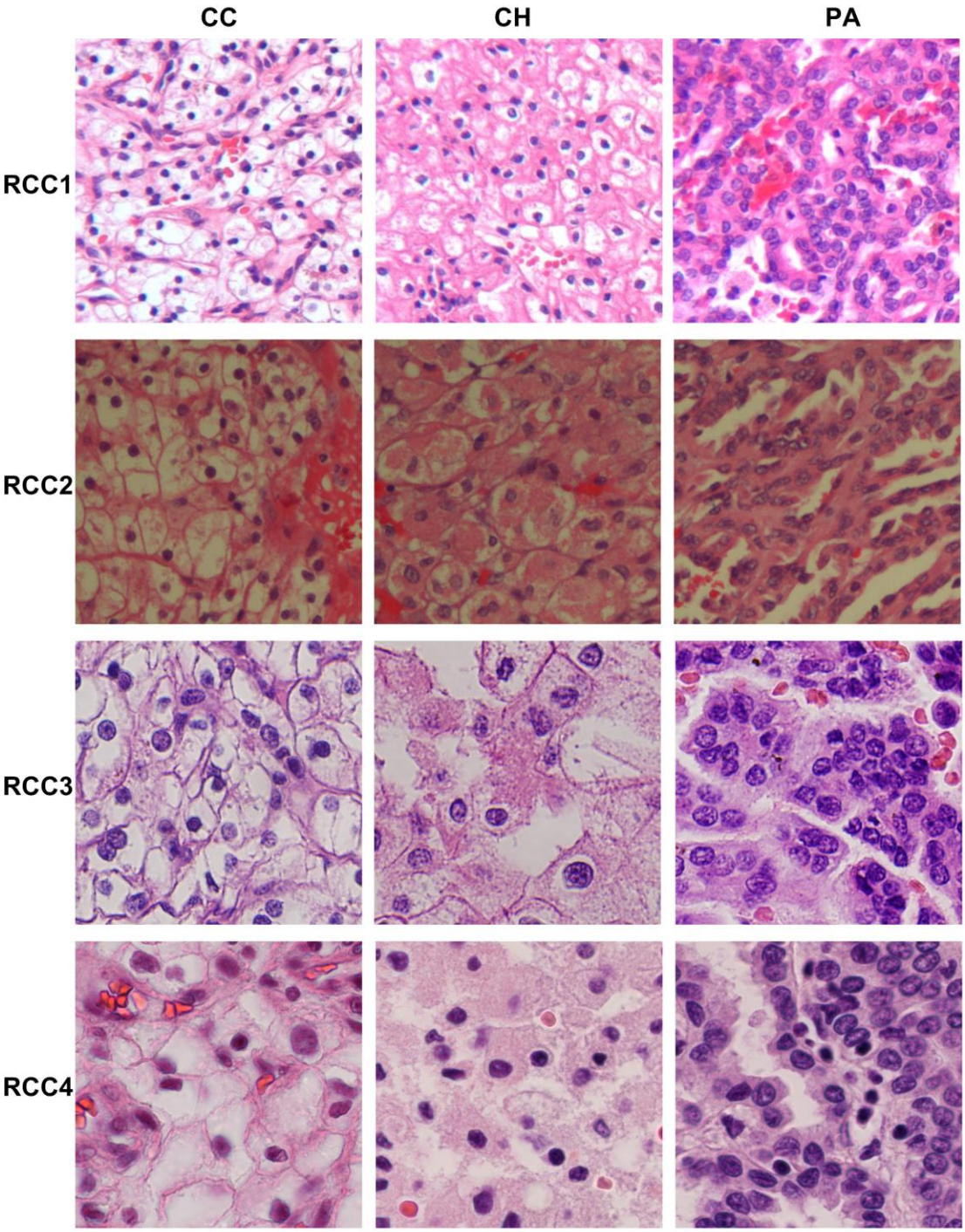

Figure 44: Image samples of three subtypes in four batches of renal cell carcinoma. Each image is a 512x512-pixel subsection of the original WSI.

Figure 45 illustrates 512x512-pixel subsections of four Fuhrman grade samples in three batches. G1 cells have small, intensely stained, nuclei with no visible nucleoli. G2 cells have finely granular chromatin, slightly textured nuclei, and inconspicuous nucleoli. G3 cells have unequivocally recognizable nucleoli. G4 have nuclear pleomorphism (varying size of nuclei), hyperchromasia (abundance of DNA, leading to darker staining) and single to multiple macronucleoli. Similar to subtypes, grades are easy to distinguish within batches based on nuclear size, shape and texture. However, grades across batches appear to be very different.

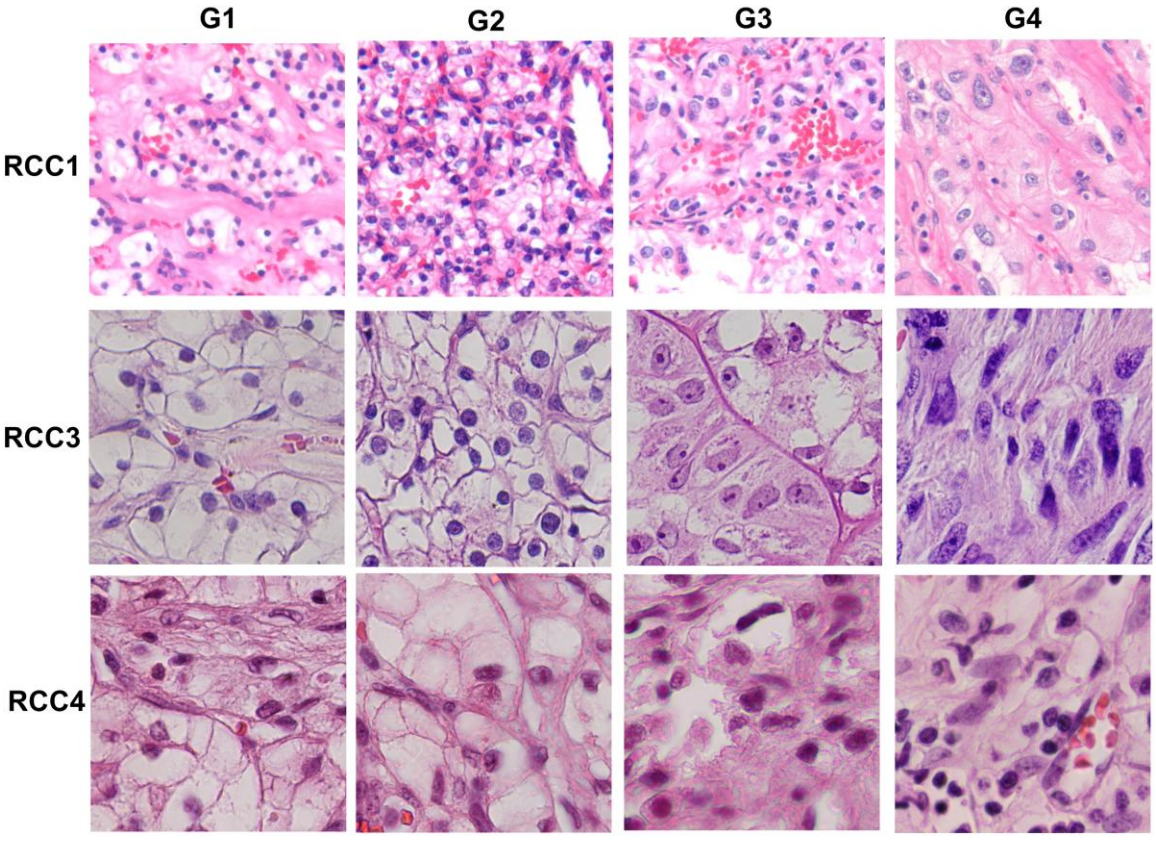

Figure 45: Image samples of four renal cell carcinoma grades in three batches. Each image is a 512x512-pixel subsection of the original WSI.

Image Feature Extraction Methods

We represent each image using the comprehensive image feature set (Table 6) except fractal features. We have excluded fractal features because of the following reasons: (1) Unlike chapter 6, we do not crop images into 512x512 subsections and fractal features can only be extracted from 512x512 subsection, and (2) In one of the normalization methods, we scale image themselves and cropped portions from scaled images would have different tissue regions. In total, we represent each image using 2663 images features and apply data mining approaches (discussed in Section II.D) to select optimal features and classification parameters. Since images in different batches are of different sizes (Table 13), we normalize the features that are affected by image size such as color or intensity histograms and nuclear count. Before extracting color features, we normalize image colors using colormap normalization (Chapter 3).

Automatic feature extraction from histopathological images is complicated because of various levels of image segmentation. In our system, we use two levels of segmentation: color segmentation to separate different stains in the image and nuclear segmentation to segment individual nuclei. We segment the stains using the methods described in Chapter 2 using OvCa images as the reference. The middle row of Figure 46 includes examples of segmented stains from different batches. After color segmentation, we obtain a binary mask for nuclear stain. However, it is still difficult to isolate individual nuclei because: (1) texture in a high-grade nucleus can break it into segments (Figure 46.G), and (2) adjacent nuclei can overlap, forming nuclear clusters (Figure 46.E and Figure 46.F). The first challenge can be addressed by merging neighboring nuclear

stain structures. However, this may lead to even more complex clusters in images with large clusters, such as the papillary subtype. Thus, we only perform a merging treatment in images with a large percent of small isolated nuclear regions, i.e., images whose nuclear mask has greater than 10% percent of regions with area less than 20 pixels. For merging, we selectively grow the regions based on their area such that larger regions are grown more than smaller regions (which may represent noise). We grow regions morphologically using radial structural element with radius, $r = A/100$ pixels, where A is the area of a region. We limit the value of r to be 3 pixels. Based on empirical analysis, we found that this treatment works better than using a morphological closing operation on all regions. Next, we segment nuclear clusters using concavity detection and ellipse fitting (Chapter 3). The bottom row of Figure 46 includes example images of segmented nuclei from different batches.

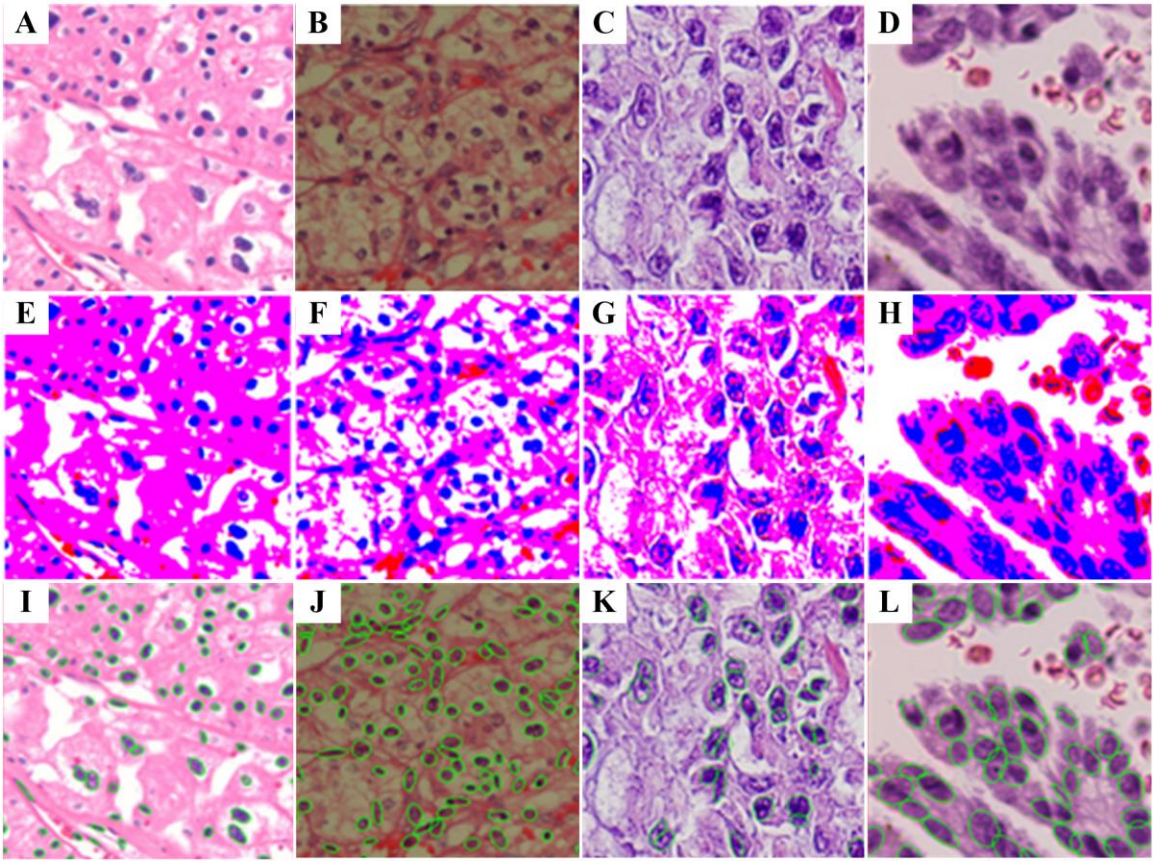

Figure 46: Segmentation results for sample renal cell carcinoma histopathological images from four data batches illustrate scale batch effects.
A, B, C, and D are 512 x512 subsections of images from RCC1, RCC2, RCC3, and RCC4, respectively. (E-F) Pseudo-colored color segmentation results, where blue, pink, white, and red represent pixels in nuclei, cytoplasm, glands, and red blood cells. (I-L) Nuclear segmentation results, where nuclei are marked using green ellipses on the original images (A-D). Note the difference in scale and nuclear size among four batches.

Normalization Methods

Scale Normalization

Upon visual inspection, we found that, besides color batch effects, which are handled using color normalization, these images differ in scale, i.e., images in batches RCC3 and RCC4 are at a higher scale compared to batches RCC1 and RCC2. Scale

differences can be calculated using the physical size of a pixel. In digital micrographs, the physical pixel size varies with factors such as microscope magnification, CCD-pixel size, and digitizing software settings. The physical pixel size is not available for RCC1, RCC2, and RCC3 datasets. Therefore, we use a model based on nuclear area to estimate scale differences and normalize all batches. Although nuclear area varies with subtype and grade in a batch, when we studied the distribution of all nuclei in a batch, we found that the distribution peaks at a specific nuclear area. Moreover, as we upscale or downscale the images, the distribution shifts right or left. Figure 47.A shows the distributions for nuclear area as RCC1 is scaled using the Lanczos (3-lobe) filter. These distributions represent all nuclei in all images in RCC1. The scaling factor, s, affects both the x and y-dimension and scales the nuclear area by a factor of s^2. Therefore, the relationship between scaling factor, s, nuclear area in a scaled batch, A_s, and nuclear area with no scaling, A_1, is given by

$$s = \sqrt{\frac{A_s}{A_1}}. \tag{39}$$

We use median area of all nuclei in the batch at scale s to quantify A_s. In Figure 47.B, we compare the empirical median nuclear area A_s of the RCC1 batch (red circles) when images are scaled for s=0.5 to 2, to the values predicted by the model using A_1 (cyan line). We can observe that values predicted by the model closely correspond to the empirical values. In Figure 47.C, we have plotted the nuclear area distribution for four batches with no scaling. These plots suggest that the four batches are at different scales and RCC1 is at the lowest scale. Therefore, we scale down RCC2, RCC3, and RCC4 with scaling factors of 0.88, 0.53, and 0.46 (calculated using (1)), respectively. In Figure 47.D, we have plotted nuclear area distribution for four batches after scale normalization. After

scale normalization, the nuclear area distributions of the four batches are more similar.

We then use scaled versions of images to extract image features for prediction.

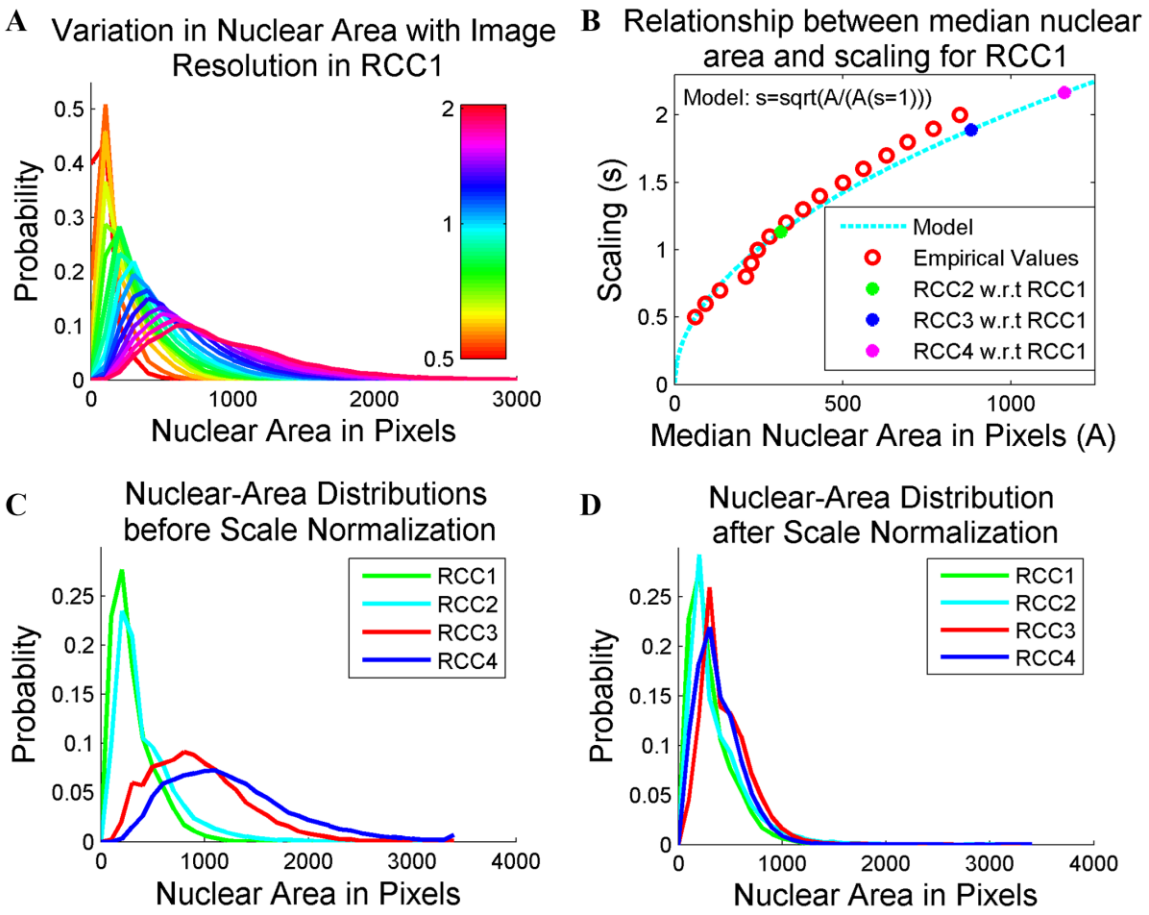

Figure 47: Scale normalization of images based on median nuclear area model. (A) Nuclear area distributions of RCC1 at different scales, (B) comparison between empirical and model-based nuclear area values for RCC1, (C) nuclear area distributions for four batches without scaling, and (D) nuclear area distributions for four batches after scale normalization. Scaling factors of other three batches with respect to RCC1 are marked in B. © 2013 IEEE

Mean Normalization

Scale normalization normalizes images to remove batch effects while all other removal methods (in this chapter), including mean, normalize features to overcome batch effects. After feature extraction, each image i in batch b is represented as a 2663-dimensional vector $Y^{i,b}$. The most prominent type of batch effect in microarray gene expression data results in differences in location (mean) and scale (spread) of gene expression values within batches. To overcome this challenge, researchers standardize the microarray data gene-wise, where for every batch, expression of a gene over all samples is adjusted such that the mean is zero and the standard deviation is one. Similarly, in mean normalization, we normalize every feature f in batch b to be within a range $[0,1]$ using the following transformation [163]:

$$\hat{Y}_f^{i,b} = \frac{1}{2}\left[\frac{\left(Y_f^{i,b} - \mu_f^b\right)}{3\sigma_f^b} + 1\right], \tag{40}$$

where μ_f^b and σ_f^b are the sample mean and standard deviation of feature f over samples in batch b, given by

$$\mu_f^b = \frac{1}{N^b}\sum_{i=0}^{N^b} Y_f^{i,b} \tag{41}$$

$$\sigma_f^b = \sqrt{\frac{1}{N^b - 1}\sum_{i=0}^{N^b}\left(Y_f^{i,b} - \mu_f^b\right)^2}, \tag{42}$$

where N^b is the number of images in batch b.

Rank Normalization

Unlike mean normalization, the rank normalization method does not assume an intrinsic normal distribution of features. Rank normalization also normalizes every feature j of a batch b to a range of $[0,1]$ [163]. We develop a rank function R_f^b, which

orders the feature values $Y_f^{1,b}, Y_f^{2,b}, \ldots, Y_f^{N_b,b}$ such that $R_f^b\left(Y_f^{m,b}\right) > R_f^b\left(Y_f^{n,b}\right)$ implies $Y_f^{m,b} > Y_f^{n,b}$. Using this rank function, we normalize the features as follows:

$$\hat{Y}_f^{i,b} = \frac{R_f^b\left(Y_f^{i,b}\right)-1}{N_b-1}. \tag{43}$$

To ensure that features are uniformly distributed in the range $[0,1]$ after rank normalization, we assign an average rank to images with the same feature value. For example, if there are n images with the same feature value, which is greater than the feature values of m images in a batch, we assign rank $m + n/2$ to all n images.

Ratio Normalization

Researchers have illustrated the usefulness of ratio-based methods for normalizing microarray expression data [27], where features are divided by the mean expression of a reference set of control samples (corresponding to each batch). In the absence of a reference set for a batch, researchers have used the mean expression of a batch in place of the control samples. However, this leads to information leakage because part of the labeled test samples is used for normalization as well. In our study, we have no reference set. Therefore, we normalize a feature f for a batch b using the median feature value for the batch M_f^b, given by

$$\hat{Y}_f^{i,b} = \frac{Y_f^{i,b}}{M_f^b}. \tag{44}$$

ComBat Normalization

Mean normalization adjusts location and scale for each feature independently. However, we can assume that batch effects are similar for most features and model batch effect-induced variance using accumulated knowledge across multiple features. Johnson et al. developed a Bayesian framework for modeling additive, γ_f^b, and multiplicative, δ_f^b, batch effects in microarray gene expression datasets [162]. They modeled batch effects using parametric (ComBatP) and non-parametric (ComBatN) priors. Similarly, we assume that $Y_f^{i,b}$, for feature f and image i in batch b, can be represented using a location and scale (L/S) model, given by

$$Y_f^{i,b} = \alpha_f + X\beta_f + \gamma_f^b + \delta_f^b \varepsilon_f^{i,b}, \tag{45}$$

where α_f is the overall feature, X is a design matrix for sample conditions, β_f are regression coefficients corresponding to X, and ε is normally distributed error with mean zero and variance σ_f^2 [162]. Johnson et al. illustrate that if the L/S model parameters are estimated using a Bayesian framework (which pools the information across features), parameters are more robust in removing batch effects. Therefore, we estimate $\hat{\alpha}_f$, $\hat{\beta}_f$, $\hat{\gamma}_f^b$, and $\hat{\delta}_f^b$ using the suggested three-step Bayes framework: (1) Calculate standardized data $Z_f^{i,b}$, which has similar overall mean and variance for all features, (2) Estimate γ_f^b and δ_f^b using parametric (ComBatP) or non-parametric (ComBatN) priors, and (3) Adjust the data for batch effects using the following equation:

$$\hat{Y}_f^{i,b} = \frac{\hat{\sigma}_f}{\hat{\gamma}_f^b}\left(Z_f^{i,b} - \hat{\gamma}_f^b\right) + \hat{\alpha}_f + X\hat{\beta}_f. \tag{46}$$

We normalize datasets using the R language implementation of ComBat provided by Johnson et al [162].

Figure 3 illustrates difference in various feature normalization methods using median-nuclear-area feature. We can observe that mean and rank normalization force different batches to have exactly same distribution of batch prevalence while ratio, combat, and combat make distributions similar but retain some original peaks.

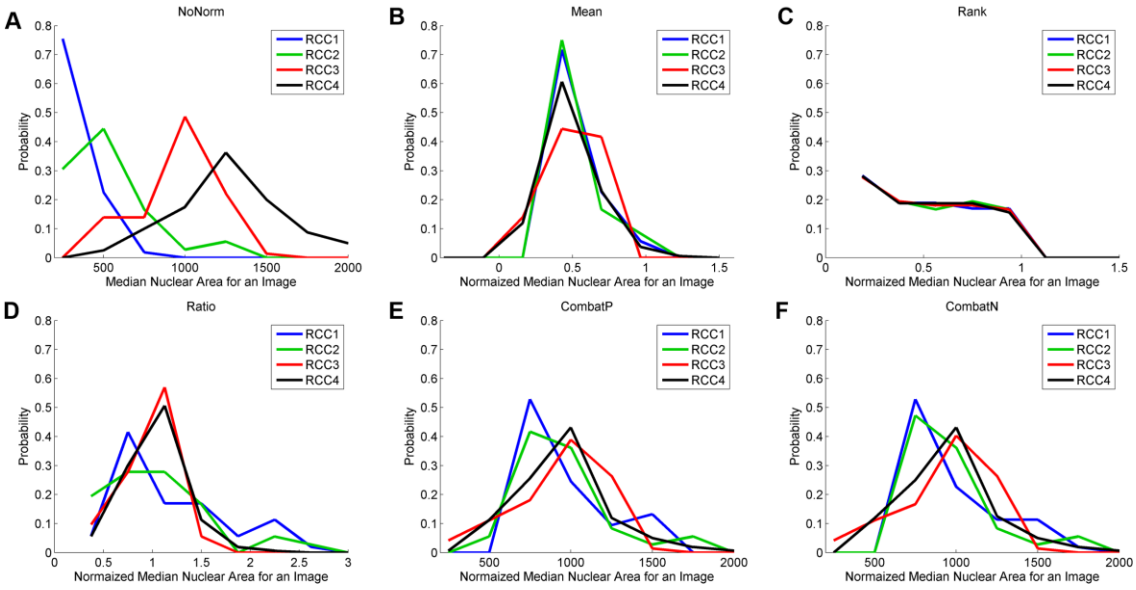

Figure 48: Normalized median nuclear (elliptical) area using different feature normalization methods.
In each subplot, curves represent feature value for all images in the study segregated by batches. Subplots capture features in different conditions: (A) without normalization and with five types of feature normalizations: (B) mean, which subtracts mean and divides by standard deviation in feature; (C) Rank, which converts feature to a rank in range of [0, 1], (D) Ratio, which divides by median value of feature; (D) CombatP, which adjust mean and scale using parametric Bayes framework; and (E) CombatN, which adjust mean and scale using non-parametric Bayes framework.

Feature Selection and Classification

We develop image-based prediction models for diagnosing renal tumor subtype and grade using data mining methods: feature selection and classification. Using the above described methods, we represent images as a large set of normalized or un-normalized

quantitative image features. However, only a few features among these features are informative for cancer prediction. Therefore, we select features using mRMR-d. mRMR is an iterative feature selection method that maximizes mutual information between features and labels while minimizing mutual information between features in the selected feature set [93]. We experiment with feature sizes in a range of 1 to 45. For classification, we develop prediction models using multiclass support vector machines (SVM) with the LIBSVM library, which returns the maximum voted label based on binary models [156]. We choose parameters for the SVM linear and radial basis kernels from the following list of cost values: $c = \{\{0.1, 0.2, \dots, 0.9\}, \{1, 2, \dots, 9\}, \{10, 20, \dots, 100\}\}$. Also, for the radial basis kernel, we select gamma from: $\gamma = \{2^2, 2^1, .., 2^{-10}\}$. We optimize all parameters—feature size, SVM cost, and SVM kernel—using 10 iterations of 3-fold CV on the train set.

Results and Discussion

Variance in Data Contributed by Batch Effects

Batch effects often influence quantitative image features to such an extent that they are a major source of variation in the data. Thus, they often overwhelm the natural segregation of images due to biological classes. To illustrate the batch effect in the data under study, we clustered image features without any normalization from all samples in four batches [162]. Figure 49 shows hierarchal clusters in the data, where the heat map highlights the feature (rows) variation across samples (columns). It can be observed that the four dominant clusters correspond to four batches.

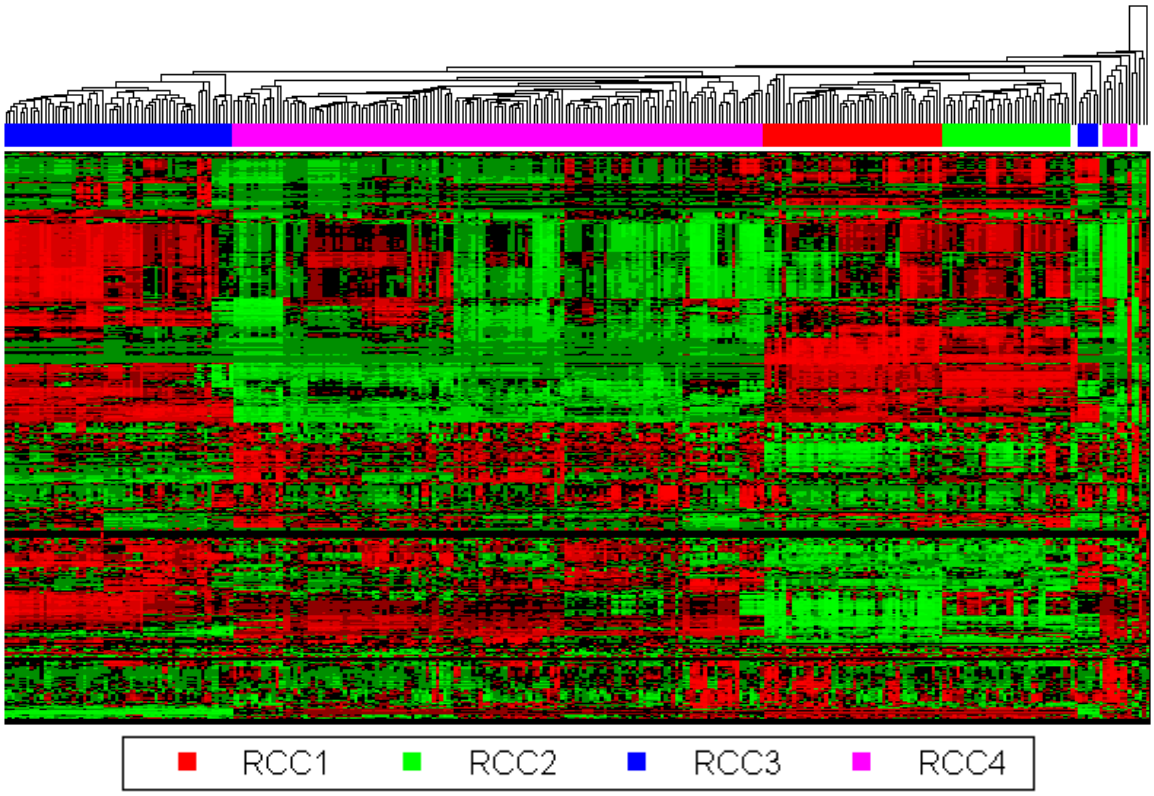

Figure 49: Unsupervised hierarchal clustering of renal cell carcinoma histopathological image features in four datasets illustrates batch effects.
Columns in the heatmap correspond to samples while rows correspond to individual features. Heatmap is normalized to show variation of a feature across different samples, where values above, at, and below the mean are represented as red, black, and greens colors. The dendrogram and horizontal colored bars above the heatmap illustrate clusters in the data. It can be observed that samples are primarily clustered based on their batch. Note: clusters without color bars contains samples from multiple batches.

We use principal variation component analysis (PVCA) to measure the variation in data contributed by the following factors: batch, grade, subtype, interaction between batch and subtype, interaction between batch and grade, and interaction between subtype and grade [26, 164]. PVCA is a useful method for calculating the proportion of variance attributable to different factors in high dimensional data. It is a combination of two

popular data analysis methods: principal component analysis (PCA) and variance component analysis (VCA). Before applying PVCA, we standardize the combined data (including three batches: RCC1, RCC3, and RCC4) using the same formula as mean normalization. We have excluded RCC2 in this chapter because it was not annotated with grades. The first step of PVCA involves reducing the dimensionality of data from 2663 features to the top few principal components (PCs) capturing a fixed portion (here 90%) of the variation (information) in the data. The second step involves applying VCA to calculate the variation in each PC contributed by each factor. For calculating the variance, VCA assumes each factor as a random effect in a linear mixed model. Variances for each PC are weighted by Eigen values of the component and averaged. Figure 50 illustrates weighted variances contributed by each factor in features with and without normalization. Variance contributed by batch is 0.365 in features without normalization, which is much higher than biological factors: grade and subtype. Variance contributed by batch is considerably reduced by all feature normalization methods: mean, rank, ratio, ComBatP, and ComBatN.

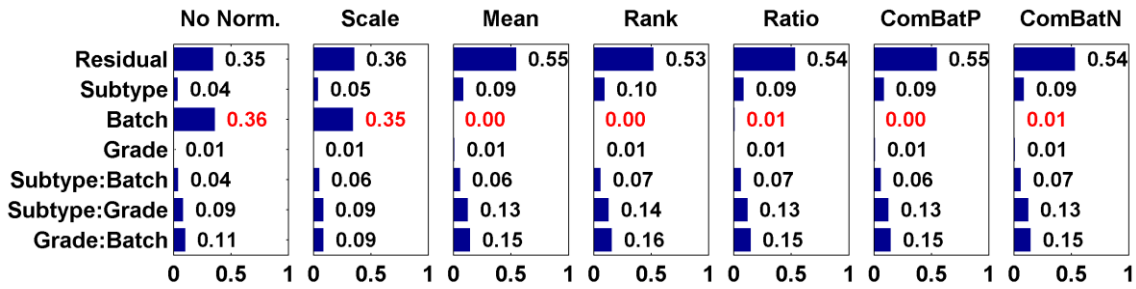

Figure 50: Principal variation component analysis on a combined dataset, including RCC1, RCC3, and RCC4, with and without normalization.
Bars indicate weighted variance contributed by various factors to the overall variation. Variance contributed by batch is significantly reduced by feature normalization methods.

We visualize the distribution of samples in the image feature space using scatter plots of component scores for the first and second principal components [27]. Figure 51 illustrates the scatter plots with and without normalization. The scatter plots show clear separation of samples from different batches (represented by four colors) in the data without normalization and scale normalization. In contrast, samples are randomly distributed in the plots for all other normalization methods.

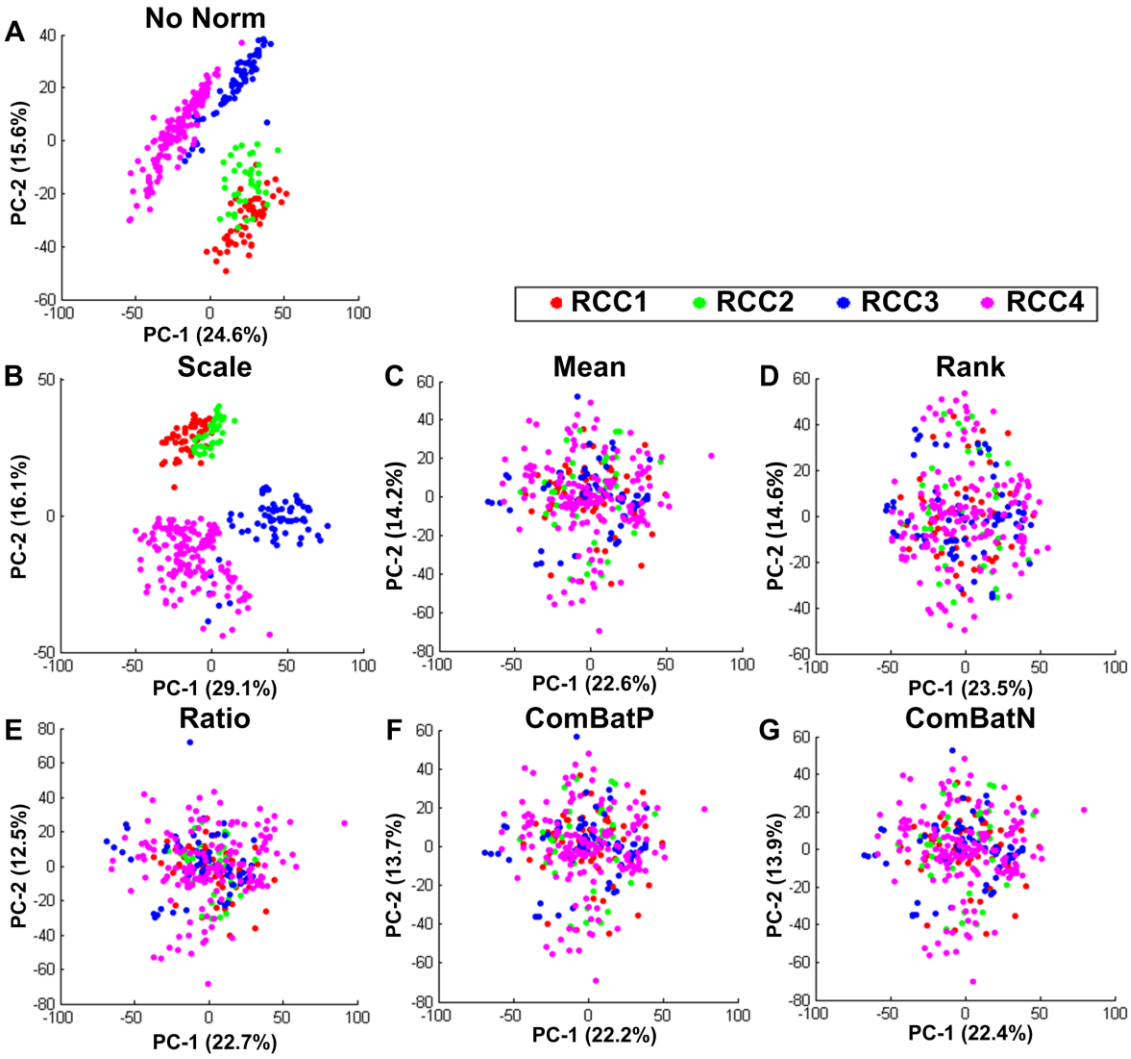

Figure 51: Scatter plots representing scores of samples for first and second principal components.
The percentage in the bracket represents percent of total variation (energy) along that component. With no normalization, batches (represented by four colors) are clearly separated. With all normalization methods except scale, separation between batches is reduced and samples seem to be randomly distributed. © 2013 IEEE

To further investigate the impact of batch effects on image features in scale-normalized data, we ranked image features that are predictive of batch using mRMR feature selection. Comparing the ranked list for scale normalized data to un-normalized data, we found the following: (1) texture properties, which are highly predictive of batches, are ranked high in both datasets and (2) shape-based properties such as median boundary fractal and Fourier error are informative for batch prediction in un-normalized data but not in scale normalized data. Therefore, besides scale, batches can differ in texture, which is not corrected by scale normalization. Differences in image formats associated with different compression methods, camera CCD, and magnification are possible causes for texture differences in the batches.

Within-Batch Prediction Performance

To verify the utility of the proposed image feature set for classifying renal grades and subtypes, we perform within-batch CV of prediction models for all batches. We observe that during CV within a batch, prediction models perform very well with classification accuracy greater than 88% for all cases except RCC1 grading (Table 15). A possible cause for the low accuracy of the grading model for RCC1 could be the small sample size, i.e., only 53 total samples are available in four classes.

Table 15: Within batch CV accuracy of multi-class renal subtyping and grading models.

	Subtype	**Grade**
RCC1	0.88	0.44
RCC2	0.88	NA
RCC3	0.93	0.99
RCC4	0.91	0.88

Cross-Batch Prediction Performance

We perform a cross-batch validation of grading and subtyping models with and without normalization. In total, with all combinations of train set, test set, and endpoint, we have 18 comparisons. Figure 52 illustrates performance accuracies for all comparisons and normalization methods. Entries in the figure are highlighted in pink or blue if the performance has increased or decreased compared to no normalization. In general, performance of prediction models are much lower compared to CV within a batch. Possible reasons for this decrease are as follows: (1) biological variance in grade or subtype is not sufficiently captured by a train batch, or (2) normalization methods are not able to completely eliminate the batch effects. However, normalization methods do improve the prediction performance of several models. When compared to prediction accuracy of features with no normalization, we observed that the ComBatN, ComBatP, mean, and rank normalization methods resulted in average performance increases of 16%, 14%, 14%, and 12%, respectively. Moreover, ComBatN resulted in the largest number of cases with prediction improvement, with 83%.

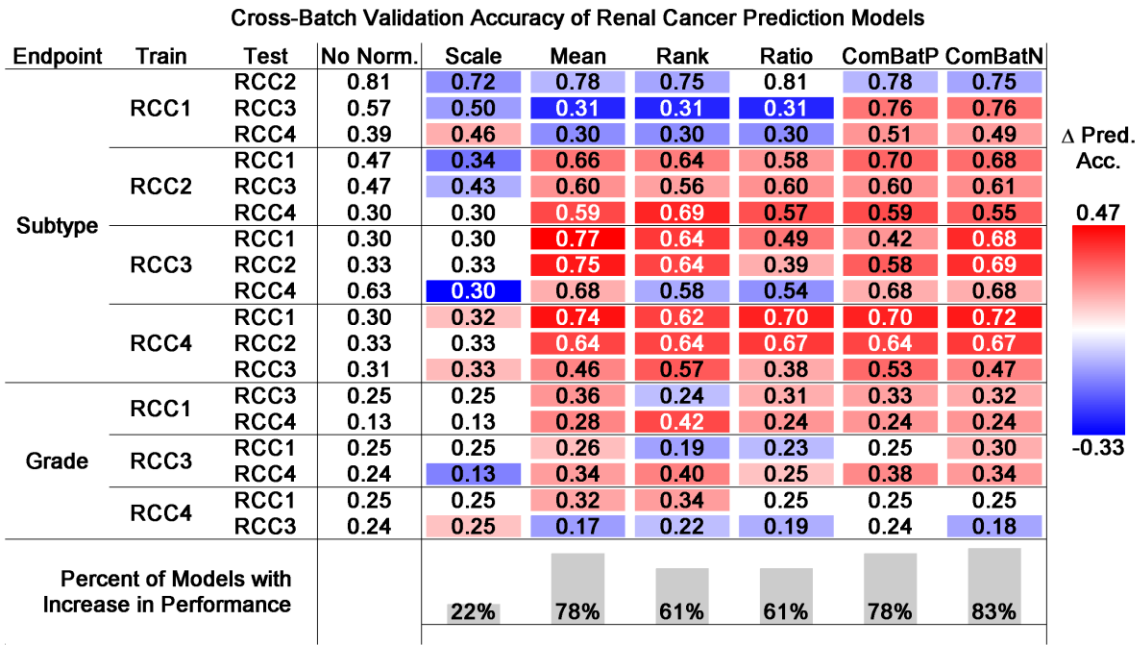

Figure 52: Cross-batch validation accuracy of renal prediction models.
The performance of normalized models is highlighted based on change in prediction accuracy compared to no normalization. Feature normalization methods, especially ComBatN, increase prediction performance for most of the models. © 2013 IEEE

Combined-Batch Prediction Performance

In a clinical setting, a CDSS is often trained with a set of images collected in batches over time. In such a scenario, batch removal methods are essential before combining the batches. We compare the prediction performance of renal endpoints with and without normalization while combining two or more batches for training (Figure 53). Similar to cross-batch validation, all normalization methods except scale normalization significantly improve performance compared to no normalization. We observed that the mean, ComBatN, rank, ComBatP, and ratio normalization resulted in average performance increases of 15%, 14%, 14%, 13%, and 11%, respectively. Moreover, ComBatN resulted in the largest number of cases with prediction improvement, with

90%. With these results, we can conclude that, even with the presence of data from multiple batches in the train set, feature selection cannot filter out features that are significantly affected by batch effects. As such, it is essential that the features are normalized before batch combination. Two prediction models in Figure 53 combine all batches. Thus, we report CV performance in these cases. The CV performances are very high and comparable to within-batch prediction performances. Thus, representation of test set samples by including similar samples (i.e., same batch) in the train set can significantly improve the performance.

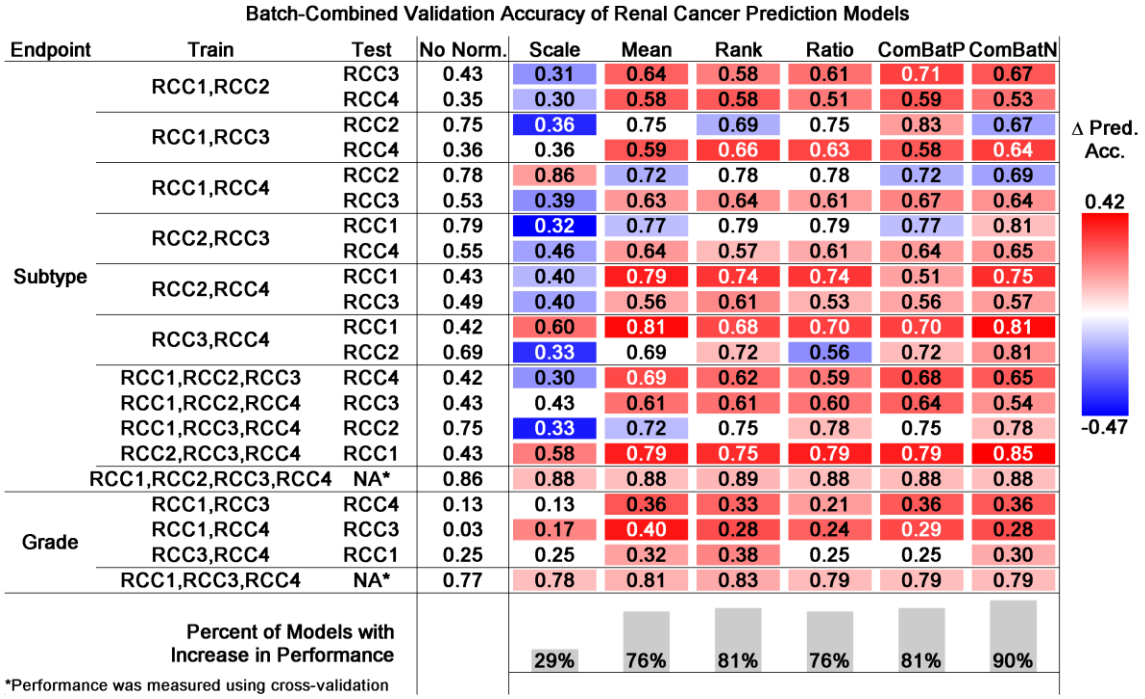

Figure 53: Accuracy of renal prediction models after batch combination. Performance of normalized models is highlighted based on change in prediction accuracy compared to no normalization. Feature normalization methods, especially ComBatN, increase prediction performance for most models. The performance significantly increases with representation of multiple batches in the train set. © 2013 IEEE

Effect of Normalization Methods on Image Integrity

Normalization methods can introduce another level of image processing in the decision making system. This processing can affect image properties and features. For instance, we use the Lanczos filter for down-sampling images in scale normalization, which may cause aliasing artifacts in the scaled images and affect texture features. Figure 54 shows down-sampling results for a benchmark Moire pattern using different scaling filters. In comparison to other filters, Laczos filter obtains a good balance between aliasing and blurring but still introduces some aliasing. Unlike scale normalization, feature normalization is performed after feature extraction and does not affect image integrity. This may be one reason for the poor performance of the scale normalization method.

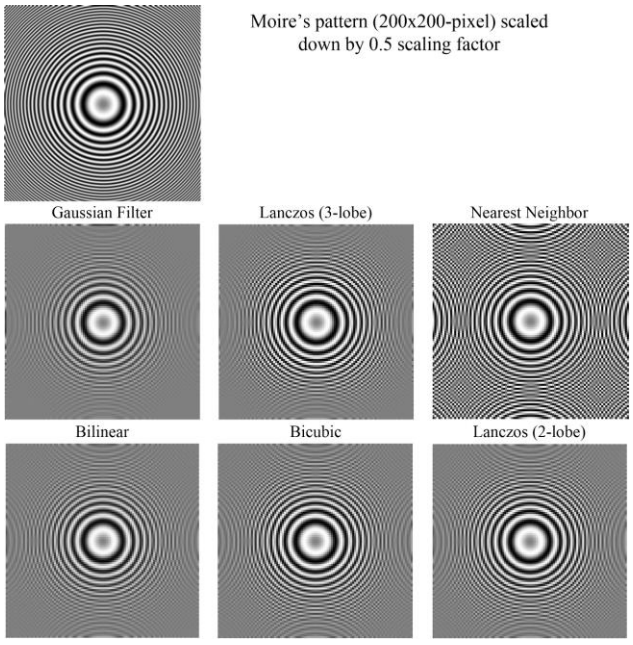

Figure 54: Down-sampling Moire's pattern image using various filters.
Gaussian, bilinear, bicubic, and Lanczos (2-lobe) filters result blurs scaled down images while nearest neighbor filter causes aliasing. Lanczos (3-lobe) achieves a balance with limited bluring and aliasing.

Conclusion

This chapter evaluated information extraction methods on subsection of histopathological WSIs, illustrated information-level batch effects, and compared multiple batch effect removal methods. Although the presence of batch effects is an important challenge for translational medicine and CDSSs that use big data, only a few researchers have investigated the impact of batch effect removal methods on histopathological image classification. Using four renal tumor image batches that have been annotated with cancer grade and subtype, we found that, compared to no normalization, ComBatN, ComBatP, mean, rank, and ratio normalization methods improve the performance of predicting renal tumor grade and subtype. In particular, ComBatN performs the best in terms of average increase in performance and total number of prediction cases with an increase in performance. Investigation of batch effects in histopathological images may become increasingly important as data repositories expand to contain valuable clinical knowledge from multiple institutions.

CHAPTER 8

TISSUEVIZ: VISUALIZATION TOOL FOR REGION-OF-INTEREST DETECTION IN WSIS

Introduction

Chapter 3 to Chapter 7 focused on the development of information extraction methods and their validation on sub-sections of histopathological WSIs. In the following chapters, we will apply these information extraction methods, including color segmentation, nuclear segmentation, and a pruned of comprehensive feature set, on WSIs and address some WSI-related informatics challenges. This chapter addresses the informatics challenge of finding region-of-interest in large WSIs. The chapter illustrates a visualization tool, called TissueViz, which facilitates the study of spatial patterns and the identification of ROIs in WSIs. The research presented in this chapter was conducted in collaboration with other researchers and most of the content is part of a published article [88].

CDSSs convert image samples into lists of quantitative image features and use pre-trained mathematical models to predict various clinical diagnoses. However, in the last decade, only a few computer-aided cancer diagnostic systems were accepted in clinical practice [2, 7, 165]. The main challenges for the translation of these systems to clinical practice include (1) the semantic gap between image descriptors and histopathology domain knowledge and (2) the noise in histopathological image samples and in subsequent extracted features. With the emergence of whole-slide digital pathology, algorithms for automating the identification of ROIs in large, heterogeneous WSIs are essential for reducing noise and improving diagnostic accuracy. We present a

visualization tool that addresses two questions pertaining to WSI analysis: (1) Do common quantitative image features form spatial patterns on WSIs? (2) Do these spatial patterns correspond to biologically relevant WSI properties such tissue necrosis or cancer grade? Visualization techniques have previously been applied to clinical histopathological images, but have only recently expanded to include WSIs.

Most existing visualization tools focus on linking image features to clinical patient information such as cancer grade, subtype and prognosis. Cruz-Roa et al. applied biclustering to simultaneously cluster image samples and code-words used in a bag-of-features description of the samples. Dendrograms from the biclustering allow users to observe the combination of code-words that represent any concept class [58]. Liu et al. used heat maps with dendrograms (TreeView software) to discover clusters among tissue microarray (TMA) samples [77]. Lessman et al. used SOMs to interpret wavelet-based features [80]. Iglesias-Rozas and Hopf used SOMs to cluster human glioblastoma samples [81]. They discovered four prominent clusters and associated these clusters with the presence of semantic histological features. Lobenhofer et al. used hierarchal clustering to visualize biologically related histopathological samples generated as part of a toxigenomics compendium study [79]. Researchers have proposed interactive information visualization tools to link multivariate bio-image features to clinical factors [76]. Besides visualizing relationships between samples and their diagnoses, some researchers have illustrated the utility of spatial pattern visualization for multivariate bio-images [166, 167]. Researchers have also developed histological image retrieval systems that annotate image blocks with semantic labels such as necrosis, glands, and

lymphocytes [62, 63]. These systems were used for analysis of rectangular histological image portions rather than WSIs.

The emergence of large data repositories such as TCGA has shifted the focus of morphological cancer analysis and digital pathology to WSIs. TCGA is a joint project by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that aims to accelerate the understanding of cancer in order to more effectively diagnose, cure and prevent cancer [3]. TCGA provides open access to high-quality genomic, proteomic and imaging data. The imaging data includes whole-slide tissue biopsy samples for several types of cancers. Recently, several researchers have used TCGA data for discovering relevant morphological properties associated with clinical diagnoses. Using ovarian serous carcinoma, Liu et al. studied the relationship between gene expression profiles, morphological properties of histological images, and chemotherapy response [168]. Soslow et al. established morphological properties associated with the BRCA1 and BRCA2 genotypes in ovarian serous carcinoma [169]. Cooper et al. discovered patient clusters in glioblastoma samples based on morphological features that have different prognoses [110]. Kong et al. associated morphological features to genomic subtypes in glioblastoma samples [170]. Chang et al. also established new subtypes in glioblastoma based on morphological features and associated them to differentially expressed genes [65]. They also developed BioSig, a system that allows the user the zoom and pan TCGAs WSIs similar to Google maps [65]. However, these systems do not have provisions for exploring quantitative image features across WSIs.

We design a visualization tool, called TissueViz, for studying visual morphological patterns across WSIs in order to address challenges related to biological interpretation of image features and to noise in heterogeneous WSIs that affects diagnostic accuracy. Using this tool, we demonstrate that (1) common histopathological image features are qualitatively associated with biologically interesting regions of WSIs, (2) sets of multiple image features can cluster sections of WSIs into biologically relevant regions, and (3) histopathology domain knowledge pertaining to ROIs can be translated to image features using supervised analysis. Our results indicate that TissueViz is useful for discovering informative image features, understanding their biological relevance, and guiding identification of ROIs in WSIs (e.g., regions with cancer cells or regions of necrosis). In the following section, we describe the TCGA image data, image processing and feature extraction, and three visualization-based analysis methods. In the results section, we show how these three visualization-based analysis methods can identify useful biological information or provide useful clinical functionality.

Materials and Methods

Datasets

We use 1301 H&E-stained WSI samples of ovarian serous carcinoma (OvCa) from 571 patients provided by TCGA [3]. For each patient, cancer grade, cancer stage and prognosis data is provided by TCGA. For each WSI, information about percent of necrosis, stromal cells, tumor cells and normal cells is also provided by TCGA. All morphological patterns discussed in this chapter and their biological interpretation was validated by a pathologist. WSIs from TCGA are available at 4 different resolutions. We

use highest and lowest (thumbnail) resolution data for our analysis. We perform quality control on the low-resolution image (Chapter 2). We crop an HxW-dimensional WSI into an MxN-dimensional matrix of 512x512-pixel non-overlapping tiles, where $M = \left\lfloor {}^{H}/_{512} \right\rfloor$ and $N = \left\lfloor {}^{W}/_{512} \right\rfloor$. Then, we select tiles with greater than 50% tissue and less than 10% tissue folds. Figure 55 shows tissue tiles, without artifacts or blank regions, selected for the OvCa WSI. We extract nine image-feature subsets, including 461 features, from each tile in the ROI of a WSI (Table 16). This feature list is a pruned form of comprehensive list used for rectangular sections (Table 6).

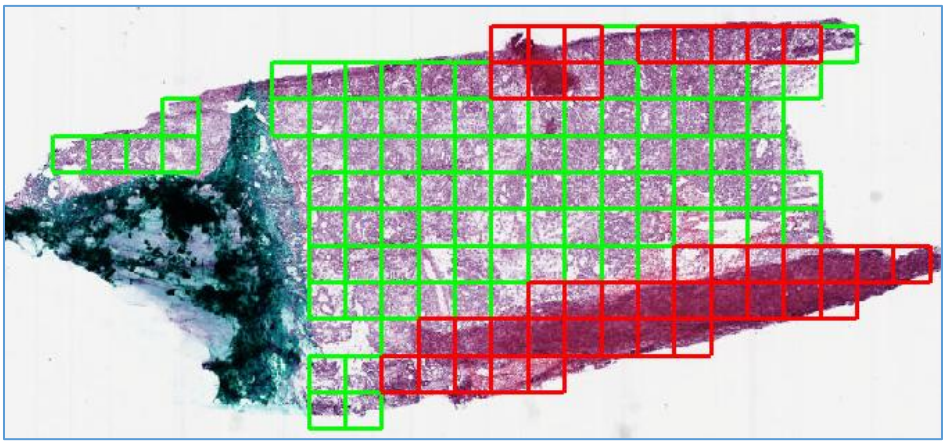

Figure 55: ROI selection in an OvCa WSI.
Squares (any color) indicate image regions with significant tissue. Red squares indicate tissue fold artifacts and green squares indicate ROI tiles. © 2012 ACM

Table 16: Pruned image feature list (461 features) used in WSI analysis.

Feature Subset	# of Features	Description
Color	73	RGB histograms, histogram statistics, and stain co-occurrence
Global texture	138	Haralick, gray-level histogram statistics, fractal, GHM multiwavelet, and Gabor
Eosinphilic-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Eosinphilic-region texture	18	Haralick and gray-level histogram statistics
No-stain-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Basophilic-object shape	51	Pixel area, elliptical area, major-minor axes lengths, eccentricity, boundary fractal, bending energy, convex hull area, solidity, perimeter, and count
Basophilic-region texture	18	Haralick and gray-level histogram statistics
Nuclear shape	26	Count, elliptical area, major-minor axes lengths, eccentricity, and cluster size
Nuclear topology	35	Delaunay triangle, Voronoi diagram, minimum spanning tree, and closeness

Visualizing Single Feature Variations

We visualize spatial patterns in WSIs formed by single image features in order to determine their biological relevance. By overlaying a heat map on a WSI thumbnail such that the heat map colors correspond to feature values, we can visually associate the feature values with morphological patterns. We map image feature values to heat map colors by discretizing the feature space into 20 equal-sized bins and use two methods for selecting the range of feature values: image-specific and dataset-specific. The image-specific feature range is calculated based on the distribution of feature values across all tiles of the WSI. Let R_n^m be the distribution of a feature, *m*, across all of the tiles in a WSI

sample, *n*. Then the dynamic range of R_n^m is given by a lower limit, L_n^m, and an upper limit, U_n^m, which are defined as follows:

$$L_n^m = \max\left\{\min\left(R_n^m\right),\ f_{0.25}\left(R_n^m\right) - 1.5 * IQD\left(R_n^m\right)\right\} \tag{47}$$

$$U_n^m = \min\left\{\max\left(R_n^m\right),\ f_{0.75}\left(R_n^m\right) + 1.5 * IQD\left(R_n^m\right)\right\} \tag{48}$$

Where function $f_p(R)$ is the p^{th} percentile of distribution *R*, and the interquartile distance, *IQD*, is given by (34).

Similarly, for the dataset-specific range L^m and U^m for each feature, *m*, are the lower and upper limits of the feature across all tiles in all WSI samples. Both methods for determining heat map range are useful. The image-specific range is useful for studying spatial patterns within individual image samples while the dataset-specific range is useful for studying spatial patterns in images with respect to the entire dataset. Figure 56 illustrates the variation of a single image feature (median Delaunay triangle area) across a WSI using the image-specific range (Figure 56.A) and dataset-specific range (Figure 56.E). This feature captures the topology/architecture of histopathology images. As such, it is lower in the region of the WSI with cohesive lymphocytes (Figure 56.A). In Figure 56, the image-specific range of the median-Delaunay-area feature is 146-567, which is the upper part of the dataset-specific range (40-454) of this feature. Therefore, the visualization in Figure 56.E is mostly red. Biologically, we observe that this WSI sample has Cribriform/pseudoendometrioid architecture (Figure 56.D). Thus, nuclei in this WSI are farther apart compared to nuclei across the whole OvCa dataset, which tends to exhibit other cohesive architectures such as papillary [169].

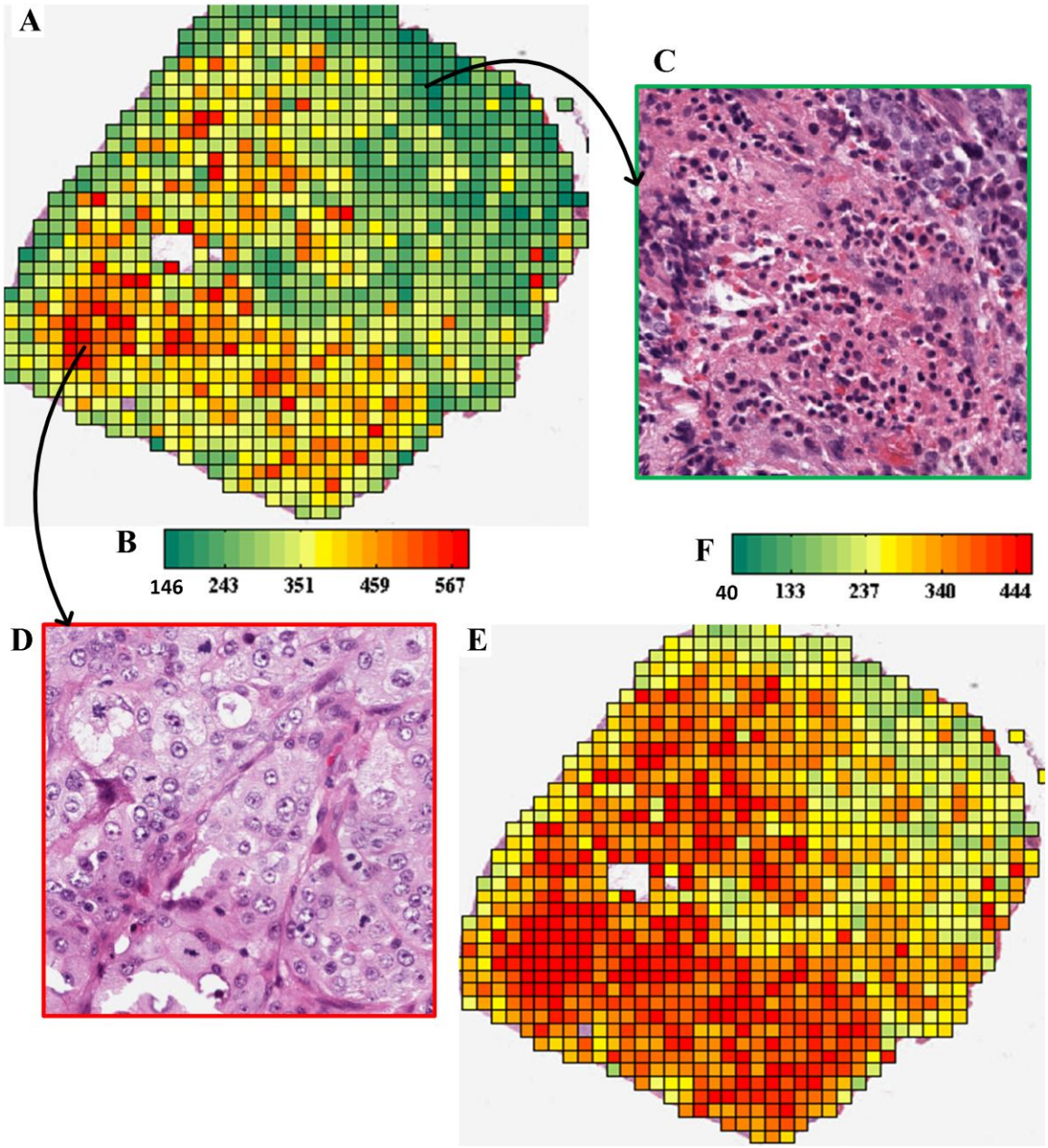

Figure 56: Visualization of variations in a single image feature (median Delaunay area, i.e. nuclear-topology feature) across the tiles of a WSI.
(A) Variations in the feature across the WSI using the image-specific range (146-577). (B) Colormap for the visualization in (A). (C) High resolution tile from the green region of the visualization in (A), indicating a low value for the feature that corresponds to more cohesive nuclear structures due to partial necrosis and lymphocytes. (D) High resolution tile from the red region of the visualization in (A). (E) Variations in the feature across the WSI using the dataset-specific range (40-454).

Unsupervised Multi-Dimensional Clustering

We study spatial patterns formed by clustering tiles in a multi-dimensional space defined by combinations of feature subsets (listed in Table 16). We use agglomerative hierarchical clustering with Ward's linkage to cluster both image features and tiles in WSIs [171]. Ward's linkage minimizes the increase of the within-cluster sum of squares as a result of merging two clusters. The increase in sum of squares by merging clusters *k* and *l* is measured by the following distance metric:

$$d(k,l) = \sqrt{\frac{2n_k n_l}{n_k + n_l}} \left\| \bar{\mathbf{a}}_k - \bar{\mathbf{a}}_l \right\|_2, \tag{49}$$

where n_k is the number of tiles in cluster *k*, $\bar{\mathbf{a}}_k$ is the centroid of cluster *k*, and $\|.\|_2$ is the Euclidian distance.

We visualize clustering by highlighting similar tiles with the same color and generating a heat map to show the grouping of both image features (rows) and tiles (columns). By studying both the WSI with highlighted tiles and the hierarchical clustering heat map, we can associate ROIs in the WSI to image feature groups. The number of clusters for visualization is a variable parameter and we observed that up to six clusters were sufficient for discovering meaningful ROIs.

Supervised Classification

Studying the variations of single image features and the clusters formed by multi-dimensional image features is useful for interpreting the biological relevance of these features. However, it is also important to link histopathology domain knowledge to image

features. In order to do this, we start by selecting two biologically different regions in the WSI, and develop a mathematical model using supervised machine learning methods to distinguish these two regions. We then use the predictive model to classify the remaining WSI tiles and visualize the result by highlighting tiles according to the probability of belonging to a selected region (Figure 57).

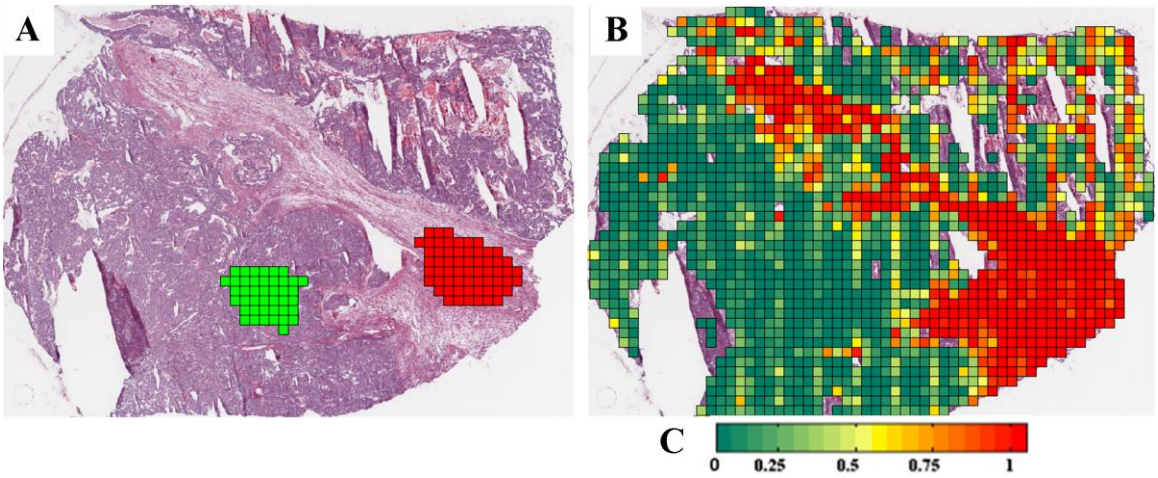

Figure 57: Visualization of Supervised Classification.
(A) Supervised selection of non-tumor (red) and tumor (green) regions from a WSI sample. (B) Based on a prediction model derived from the supervised tile selection, remaining tiles are highlighted with the probability of being in the non-tumor class. (C) Color map for the visualization in (B). © 2012 ACM

This visualization is useful for (1) identifying image features associated with biologically distinct WSI regions and, based on these features, for (2) identifying image tiles that are similar to these regions. We use the mRMR-q feature selection method for ranking features [93]. We then use the top features to develop a classification model using the linear-SVM classifier (LibSVM) [156]. We optimize the number of features used in the classification model using 5-fold, 10-iterations of CV accuracy. The model

may be incrementally trained by iteratively adding training samples (from different WSIs) to provide the classifier with more challenging training samples [172]. In Figure 57, we illustrate the selection of tumor (green) and non-tumor (red) regions of a WSI and show the result of classifying the remaining image tiles by visualizing the probability of each tile belonging to the non-tumor region (Figure 57.B, red indicates higher probability).

Graphical Tool Design

Results and Discussion

We studied OvCa WSIs using three visualization methods and found several biologically interpretable morphological patterns. In this section, we discuss an example pattern for each visualization method.

Pattern based on Average Basophilic-Object Eccentricity in Grade-3 OvCa WSIs

We studied spatial patterns captured using average basophilic-object (i.e., nuclear object) eccentricity (*ec*), a morphological shape property. Using the dataset-specific range of values for the visualization heat map, we observed that image tiles in the stroma/necrosis region have high *ec* while tiles with lymphocytes have low *ec*. This is due to the spindle-like nuclei of stroma/necrosis regions and the circular disc-like structures of lymphocytes [173]. Nuclei in tumor cells have intermediate eccentricity. Figure 58 is an illustration of spatial patterns formed by *ec* for two WSI samples. Both

samples are from patients with grade-3 OvCa but have different morphology. The first sample (Figure 58.A-E) is reported to have 3% lymphocytes, 15% necrosis, 0% stroma, and 95% tumor, while the second sample (Figure 58.G-K) is reported to have 60% lymphocytes, 0% necrosis, 5% stroma, and 95% tumor.

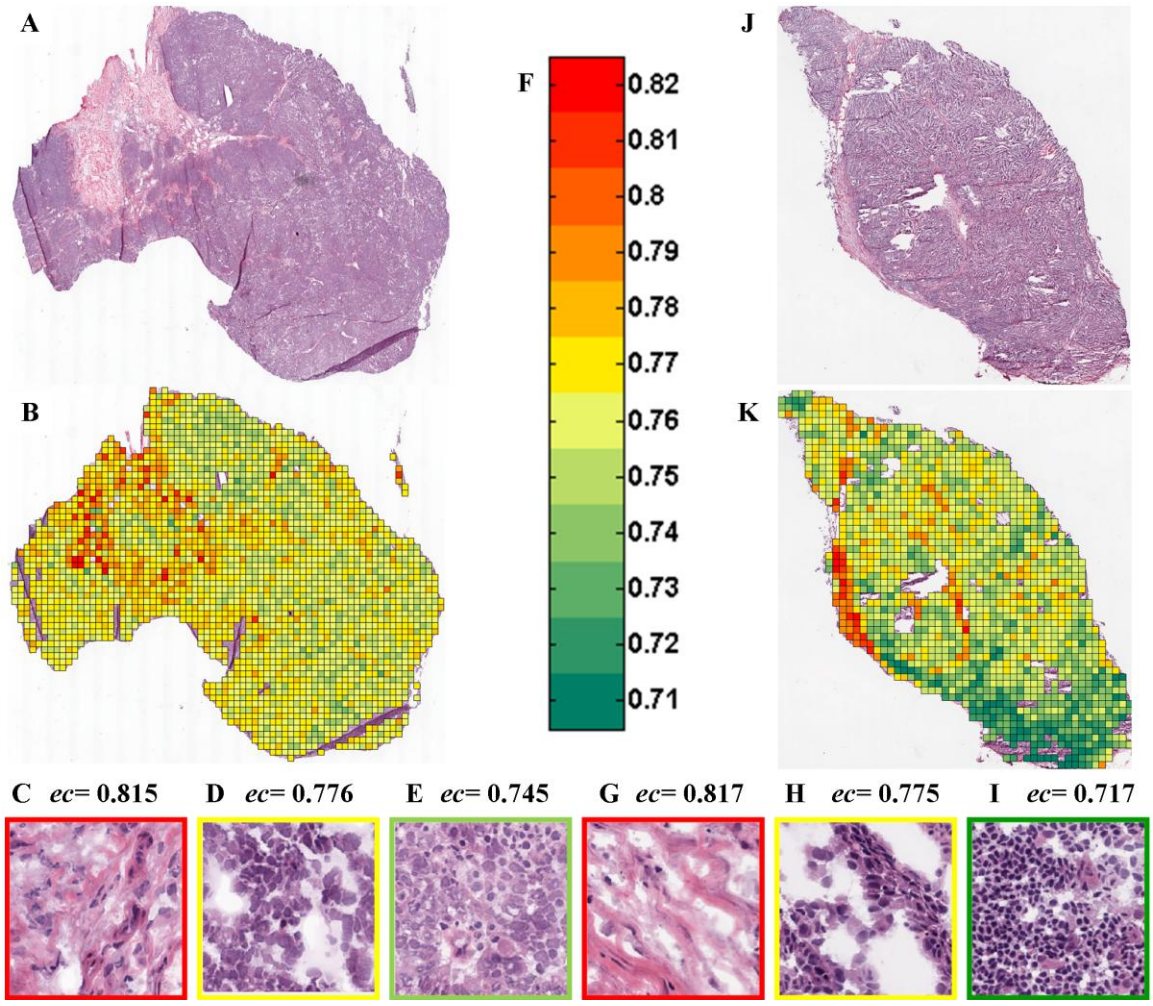

Figure 58: Variations in average nuclear eccentricity (*ec*) across two grade-3 OvCa WSI samples.

High *ec* (red) corresponds to stroma or necrotic regions with elongated nuclei, low *ec* (green) corresponds to lymphocyte-infiltrated tumor regions with circular lymphocytes, and medium *ec* (yellow) corresponds to tumor regions. (A) WSI sample with 3% lymphocytes, 15% necrosis, 0% stroma and 95% tumor. (B) WSI sample with 60% lymphocytes, 0% necrosis, 5% stroma and 95% tumor. (F) Colormap corresponding to *ec* values. (C-I) High-resolution image tiles from sample A (C-E) and sample B (G-I). Numbers above the high resolution tiles indicate the feature value. These tiles are also outlined with a color corresponding to the highlighted tile in (B) and (K).

We observed that: (1) due to the higher percentage of lymphocytic infiltration, the visualization of the second sample contains more green tiles (Figure 58.K), (2) tiles in stroma/necrosis regions are highlighted with shades of red, (3) high resolution images confirm that tiles with low, medium and high *ec* are from stroma/necrosis, tumor and lymphocytic-infiltrated tumor regions, respectively. Lymphocytic infiltration is considered useful for predicting patient prognosis [173] and has been correlated to the BRCA1 genotype in OvCa [169]**.**

Pattern based on Color and Basophilic-Object Shape Features in Grade-3 OvCa WSI

We studied spatial patterns formed by the morphological color and basophilic-object shape feature subsets that include 124 features. Using Ward's linkage with agglomerative clustering (as described in section 2.3.2), we illustrate the spatial patterns formed by three clusters of tiles in a grade-3 sample reported to have 15% necrosis, 15% stroma, and 70% tumor (Figure 59.A and Figure 59.B). Figure 59.D-F are high resolution tiles from each of the three clusters, capturing tumor (yellow), necrosis (magenta), and stroma (cyan) regions, respectively. Out of 1174 tiles in the sample, 603, 316, and 255 tiles were clustered into each of these clusters (C1, C2, and C3), respectively. The heatmap (Figure 59.C) shows that there are several correlated features in these subsets that separate these clusters. We use the mRMR-q feature selection method to find the top five most informative, uncorrelated features in these subsets that will separate C1, C2 and C3.

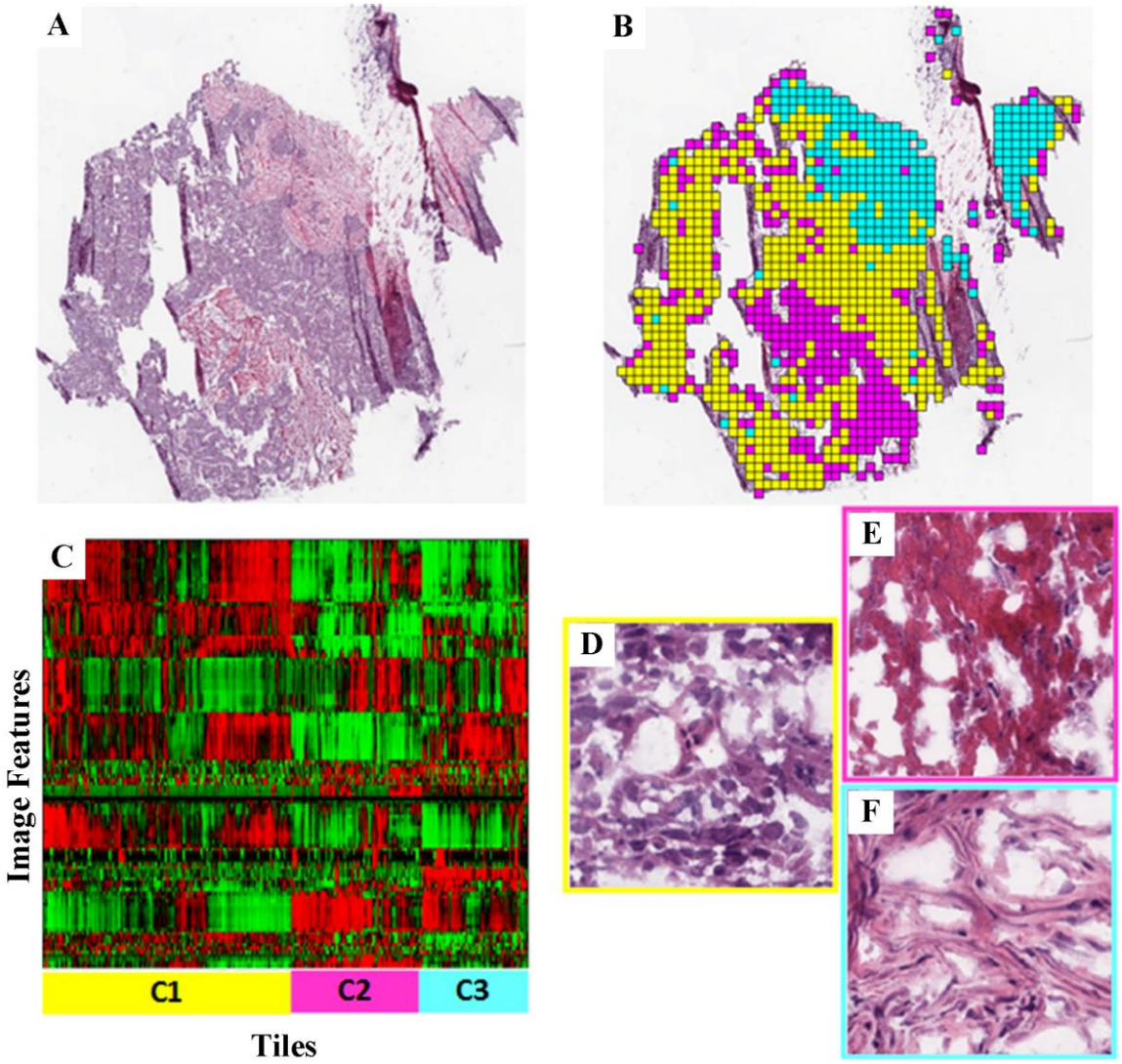

Figure 59: Three distinct tile clusters (B) in a grade-3 ovarian serous cystadenocarcinoma WSI sample (A) with 15% necrosis, 15% stroma and 70% tumor.
Clusters were found in the feature space defined by color and nuclear-object-shape features (a total of 124 features). (C) The heatmap illustrates the variations in features (row) across various image tiles (column). Red and green values in the heatmap correspond to values above and below the mean feature value. Out of 1174 tiles in the sample, 603, 316, and 255 tiles were distributed into clusters C1, C2, and C3, respectively. (D-F) High-resolution image tiles belonging to the three clusters. Clusters C1, C2 and C3 correspond to regions of tumor, necrosis and stroma, respectively. © 2012 ACM

In Table 17, we report these features (in the order of preference) with their average value and standard deviation for each cluster. These features can be easily visually associated with morphological properties of tumor, necrosis, and stroma. Due to the red appearance of stroma/necrosis regions compared to tumor regions, color features are important. Moreover, as discussed in the previous section, stroma/necrosis regions have spindle-like nuclei. Therefore, C2 and C3 have larger standard deviation in major-axis length and larger maximum eccentricity of basophilic structures. Due to the more elliptical nuclei in tumor regions, the mean boundary-fractal dimension is higher than that of stroma and necrosis. Therefore, using this visualization we can discover morphological patterns in WSI based on multiple features.

Table 17: Top five differentially expressed features in three hierarchal clusters corresponding to necrosis, stroma, and tumor regions in OvCa.

Feature Name	**C1**	**C2**	**C3**
Percent of pixels in R histogram bin (80-96)	0.025 ± 0.004	0.017 ± 0.004	0.015 ± 0.004
Percent of pixels in G histogram bin (127-143)	0.052 ± 0.006	0.040 ± 0.006	0.058 ± 0.007
Std in major-axis length of basophilic structures	6.669 ± 2.049	8.048± 2.402	9.174± 2.505
Maximum eccentricity of basophilic structures	0.975 ± 0.009	0.971 ± 0.011	0.987 ± 0.007
Mean boundary-fractal dimension of basophilic structures	1.505 ± 0.010	1.486 ± 0.020	1.458 ± 0.019

Identification of Tumor and Non-Tumor ROIs in OvCa WSIs

We developed a classification model that separates tumor and non-tumor (including stroma and necrosis) in OvCa WSIs. We selected this model because pathologists generally focus on tumor regions, excluding necrotic and stroma regions, while making histological diagnoses. Therefore, pathologists can use this model to facilitate selection of ROIs in WSIs. For this model, we performed training using tumor and non-tumor regions from 15 WSIs. On each WSI, a pathologist marked tiles in both non-tumor and tumor regions (Figure 57.A). We use these annotated tiles as ground truth. In total, 2098 tiles were used for training, among which 638 and 1460 tiles were from non-tumor and tumor regions, respectively. We performed 5-fold, 10-iterations of CV to select an appropriate feature size (from 1 to 20 top features out of 461 features). Based on average classification accuracy, we selected 17 as the optimal feature size. Table 18 lists the top 5 features.

Table 18: Top five informative features for tumor/non-tumor classification in OvCa.

Feature Name	**Tumor**	**Non-tumor**
Adjacency of basophilic stained pixels per unit area	0.400 ± 0.080	0.277 ± 0.091
Energy of Gabor filter response with frequency, $f = \sqrt{2}/64$ cycles per pixel, and orientation, $\theta = \pi/2$ radians.	57.103 ± 7.564	51.662 ± 10.845
Percent of pixels in B histogram bin (191-207)	0.059 ± 0.013	0.053 ± 0.023
Eosin stained pixels' mean intensity	0.523 ± 0.016	0.536 ± 0.021
Mean boundary fractal of basophilic structures	1.491 ± 0.014	1.472 ± 0.018

Similar to Table 17, there are some color and basophilic-object shape properties that were useful for classifying tumor and non-tumor regions. In addition, energy of

Gabor filter response is higher for tumor regions than non-tumor regions. Gabor filter response is high if there are edges in the image in the $\theta + \pi/2$ direction. Higher Gabor filter response in the tumor region is mostly due to stronger edges formed in the grayscale image by basophilic-structures compared to eosinphilic-structures. Also, mean intensity of eosin stained pixels is slightly higher (brighter/whiter) in non-tumor compared to tumor. We validate this classification model on a separate test set of 100 OvCa WSIs. Again, tiles in the tumor and non-tumor regions of these images were marked by a pathologist. In total, there were 29424 tiles in the testing dataset with 17181 tumor and 12243 non-tumor tiles. Overall classification accuracy on the testing dataset was 90%. In Table 19, we have provided a confusion matrix for the classification model performance on the test set.

Table 19: Confusion matrix for tumor/non-tumor model for OvCa.

		Predicted label	
		Tumor	**Non-tumor**
Actual Label	**Tumor**	87%	13%
	Non-tumor	6%	94%

Accuracy= 90%, Number of tumor tiles=17181, Number of non-tumor tiles=12243

Most of the error is due to false negatives, i.e., our system classifies borderline cases as non-tumor rather than tumor. From the point of view of highlighting useful regions for diagnosis, false negatives are more favorable than false positives.

We also apply this model to all tumor samples in the OvCa TCGA dataset (including the training samples) that have reported percentage of tumor cells (in total 1094 WSIs). Then we calculated the approximate percentage of tumor cells in the WSI using percent of tiles in the tumor region. We included tiles in the tissue fold regions of WSIs on the assumption that reported percentages of tumor cells are with respect to the entire samples. We found that the predicted percentage of tumor cells was correlated with TCGA-reported tumor cell percentage (Pearson's correlation coefficient = 0.4293, p-value = 2.8134e-50). Figure 60 is a scatter plot of reported percentage of tumor cells vs. predicted percentage of tiles in the tumor region. Data points are colored based on the density of points in the region. A large number of points are on the diagonal (high correlation). The points that are not correlated are mostly below the diagonal, indicating that our system tends to classify more tiles into the non-tumor region compared to what is reported by TCGA.

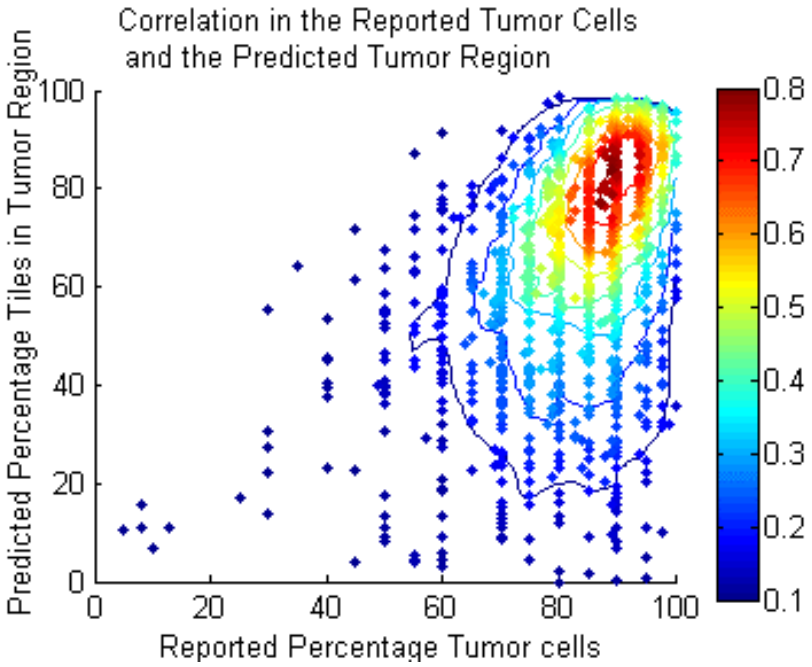

Figure 60: Correlation between reported percentage of tumor cells and predicted percentage of tiles in tumor region using a tumor vs. non-tumor classification model for OvCa.
This plot includes all WSI TCGA samples that have percentage of tumor cell information available (1094 images). Tissue fold artifacts were included when calculating the percentage of tumor regions. Pearson's correlation coefficient = 0.4293 (p-value = 2.8134e-50). The color represents the density of data points per unit of area.

Identification of Grade-4 ROIs in KiCa WSIs

Similar to the tumor vs. non-tumor classification model for OvCa, we also developed a tumor vs. non-tumor classification model for KiCa and selected tumor ROIs in WSIs of KiCa patients. Thereafter, we developed a classification model for predicting grade-4 regions among tumor regions. For training the grade-4 classification model, we selected 2815 tumor tiles from five grade-1 patient samples and 1266 tiles from five grade-4 patient samples. We use the mRMR-q (minimum-redundancy maximum-relevance) feature selection method for ranking features [93]. We then use the top features to develop a classification model using the linear-SVM classifier (LibSVM) [156]. Using 5-fold, 10-iterations of cross-validation accuracy, we found 17 as optimal feature size in range of 1 to 20.

Figure 61 illustrates prediction of the optimized model in WSIs from grade 1, grade 2, grade 3, and grade 4 patients. The color maps in Figure 61.B, E, H, and K illustrate the probability of each tile belonging to grade-4 class (red indicates higher probability). We can observe that WSIs from patients with higher grade have larger number of red tiles, i.e. more tiles are likely to be in grade 4 class.

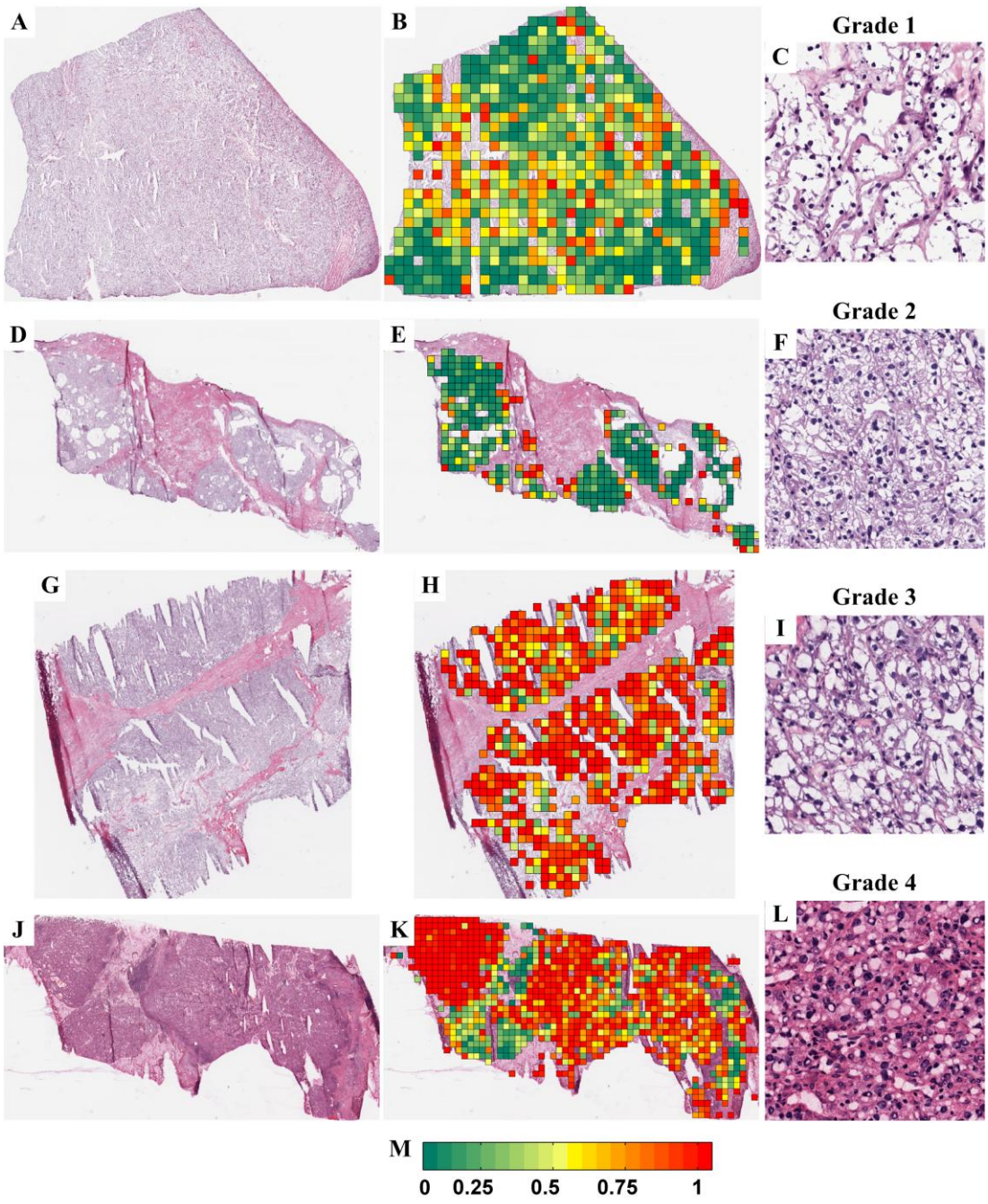

Figure 61: Prediction of Grade-4 regions in WSIs of KiCa patients.
A binary classification model was developed selection tiles from grade 1 and grade 5 samples in supervised classification mode. (A, D, G, and J) Thumbnails of WSIs from grade 1, grade 2, grade 3, and grade 4 patients, respectively. (B, E, H, and K) Thumbnails highlighted with the probability of being in the grade 4 class. Among all tumor tiles, 26% tiles were predicted as grade 4 in grade 1 sample; 20% in grade 2 sample; 85% in grade 3 sample; and 75% in grade 4 sample. (M) Color map for the visualization. (C, F, I, and L) high-resolution tiles from four samples in (A, D, G, and J).

To further validate the performance of this model, we applied the model on a larger dataset including 893 WSIs of tumor samples from 443 KiCa patients (including the training samples). We found that the predicted percentage of tumor tiles that was classified as grade 4 is correlated with patient grade reported by TCGA (Pearson's correlation coefficient = 0.4138, p-value = 0.3501e-20). Figure 62 illustrates the percent of tiles predicted as grade 4 in tumor samples of grade 1 to grade 4 KiCa patients. For each grade, a boxplot illustrates median value (red bar), interquartile range (thick blue bar), extreme range (thin blue whiskers), and outliers (red "+"). We found that the percent of tiles predicted as grade 4 is statistically different in samples of different grade patients (Table 20). Although median value for percent of tiles predicted as grade 4 increases with increasing grade, it is not zero for lower grade. Therefore, unlike pathologists, who diagnose patient based on the highest grade regions present in a tissue sample, automated systems have to allow for some high-grade predictions even in low grade samples.

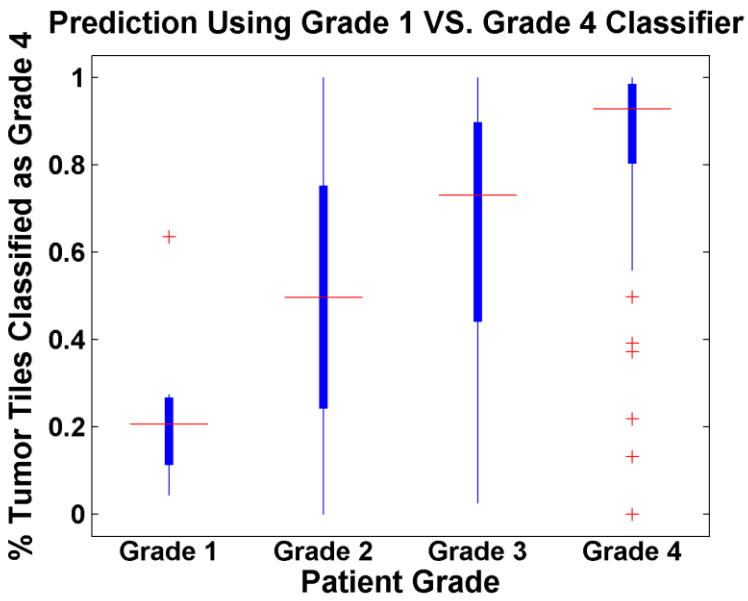

Figure 62: Percent of tumor tiles predicted as grade 4 in WSIs from patients with different grade KiCa.

Table 20: P-values for pair-wise two-sided Ranksum test between percent grade 4 prediction in samples from patients with different grade KiCa

	Grade 1	**Grade 2**	**Grade 3**	**Grade 4**
Grade 1	1	0.0104	0.0003	7.2e-05
Grade 2	0.0104	1	1.8e-07	3.9e-14
Grade 3	0.0003	1.8e-07	1	1.5e-06
Grade 4	7.2e-05	3.9e-14	1.5e-06	1

Pattern based on Nuclear Shape and Topology Features in Grade-4 KiCa WSI

We studied spatial patterns formed by the morphological nuclear shape and topology feature subsets that include 61 features. Using Ward's linkage with agglomerative clustering, we illustrate the spatial patterns formed by four clusters of tumor tiles (non-tumor tiles were eliminated using tumor vs. non-tumor classification model) in a grade-4 sample (Figure 63.A and Figure 63.C). Figure 63.B illustrates probability of tile being a grade-4 tile using a grade 1 versus grade 4 model. Among the four clusters, one cluster (brown) is an outlier with only one tile. Figure 63.E-G are high resolution tiles from remaining three clusters, capturing low (cyan), medium (magenta) and high (yellow) grade. The heatmap (Figure 63.D) shows image features for tiles in yellow cluster are very different compared to image features of tiles in other clusters.

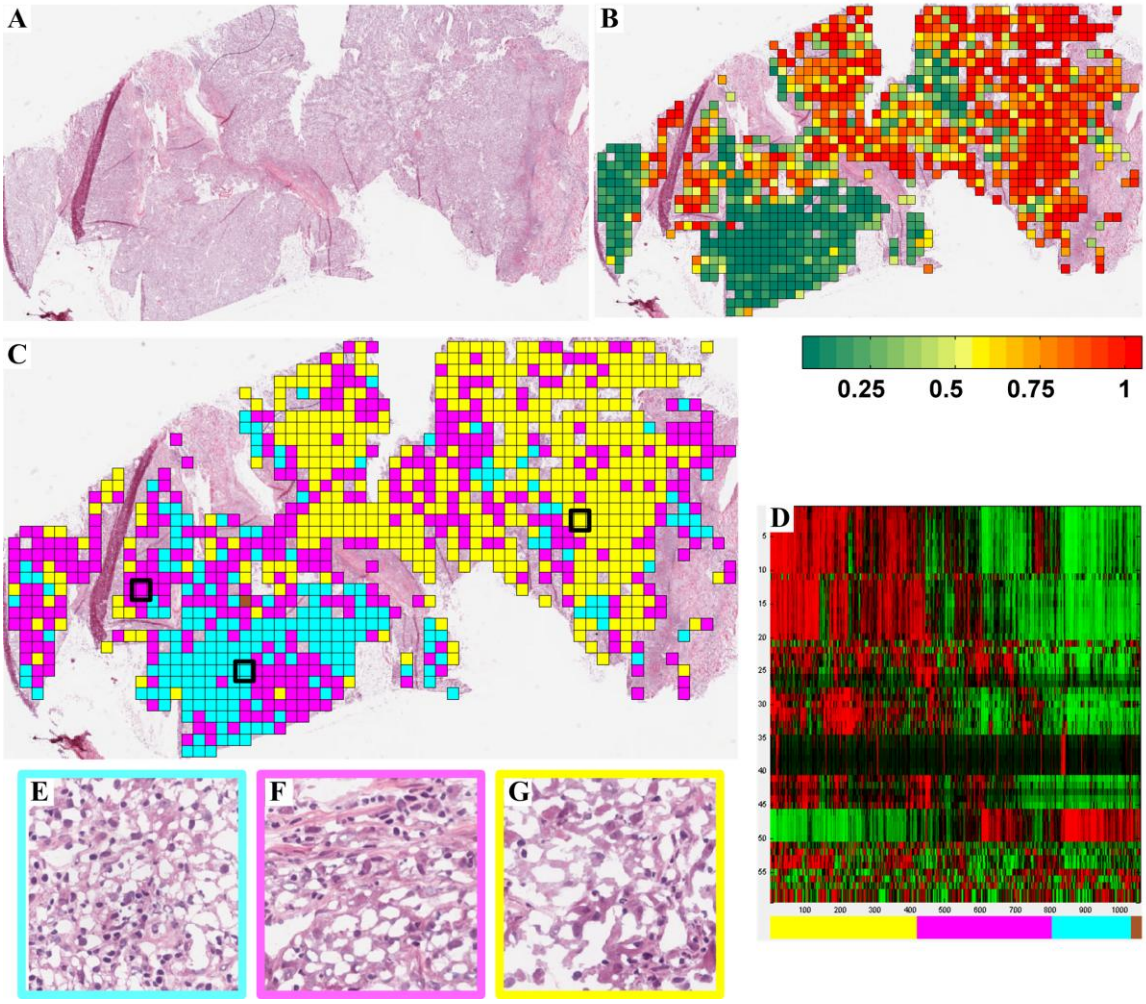

Figure 63: Clustering of tumor tiles in a WSI of a patient with grade-4 KiCa.
(A) Thumbnail of original WSI. (B) WSI painted with probability of being in the grade 4 class. (C) WSI painted based on tile's cluster, where among four clusters brown is an outlier with one tile, yellow corresponds to high grade, cyan corresponds to low grade, and magenta corresponds to intermediate grade. (D) The heatmap illustrates the variations in features (row) across various image tiles (column). Red and green values in the heatmap correspond to values above and below the mean feature value. Most features in yellow cluster appear very different than other clusters. (E, F, and G) high-resolution tiles in three clusters.

Conclusion

We have designed a visualization tool, called TissuViz, for studying morphological patterns across histopathological WSIs. This visualization tool highlights variations across regions in WSIs based on one or more quantitative image features. The visualization can be used to study patterns formed by features (single feature variations and unsupervised multi-dimensional clustering) or it can be used to identify region-of-interest (ROIs) in large WSIs (supervised classification). Although we have used ovarian serous cystadenocarcinoma (OvCa) WSIs provided by TCGA, our methods can be applied to any WSI dataset.

CHAPTER 9

QUANTIZED REPRESENTATION OF WSIS FOR ENHANCED DECISION MAKING

Introduction

This chapter addresses the informatics challenge of developing high-level representations for WSIs that can model pathologists' knowledge. High-level representations are essential to tackle biological variation in WSIs while making diagnostic decisions. The chapter also validates these representations for two applications: (1) diagnosis of histopathology-based endpoints such as subtype and grade and (2) prediction of clinical endpoints such as metastasis, stage, lymphnode spread, and survival.

CDSSs provide an objective and efficient means for diagnosing cancer using histopathological images [5]. Although many current systems use manually selected rectangular sections of biopsy slide images, the advent of whole-slide images (WSIs) and their availability in public data repositories such as TCGA has shifted the focus of research [3, 24, 65, 73, 110]. WSIs provide a holistic histopathological snapshot of a patient. Because of their size, WSIs are more likely to capture all relevant histopathological characteristics of a disease. Moreover, effective systems that use WSIs would be less likely to be biased by subjective ROI selection [174]. Thus, it is important to develop effective systems that can objectively and automatically diagnose disease using WSIs.

Image processing methods developed for analyzing only limited rectangular portions of histopathological images can be used for analyzing WSIs. However, WSIs

pose unique challenges: (1) the images are large, hindering the computational feasibility of image processing, and (2) the magnitude of biological and non-biological variations adversely affects image descriptors. To handle large images, researchers have proposed parallel, piece-wise processing of WSIs by cropping them into equal-sized, non-overlapping tiles [24, 65, 73, 110]. Features from individual tiles of a WSI are then combined to represent the patient [65, 110]. Alternatively, each tile is independently classified and all classification decisions are combined to obtain the final patient diagnosis [24, 73]. To handle experimental variations, researchers have suggested quality control methods that remove artifacts such as tissue folds [17]. Finally, to handle biological variation, researchers have suggested methods for automated ROI selection, which can be dependent on the disease, or decision endpoint. For instance, pathologists focus on high-grade tumor regions of WSIs when the endpoint is cancer grade [45]. Therefore, researchers have proposed cancer-dependent ROIs for grading such as regions with high Gleason score for prostate cancer [73] and regions with high mitotic activity for breast cancer [75]. Unlike cancer grading, prediction endpoints such as patient survival and disease stage are not associated with ROIs because histopathological images are not normally used to predict these endpoints. However, there is evidence that histopathological image features are predictive of these endpoints [57, 110, 175]. Thus, automated analysis of WSIs may also reveal novel insights into the biology of disease.

We propose a novel strategy to handle biological variation in WSIs. Instead of finding optimal ROIs for each cancer endpoint, we propose to represent WSIs using a profile of different types of biological regions. To implement this strategy, we propose

quantization methods for image representation, which quantizes the image feature space and quantifies the percentage of tiles in each quantized space. Tiles are selected from quality-controlled, tumor regions of WSIs (Figure 64). We propose three quantization methods: (1) Univariate quantization, which quantizes each feature into equal-sized bins, (2) Multivariate quantization, which quantizes the complete image feature space into unsupervised clusters, (3) Multivariate subset quantization, which quantizes subsets of similar features into unsupervised clusters. We use WSIs of patients with kidney and ovarian carcinoma obtained from TCGA [3]. We examine the effect of different quantization methods on the prediction of eight clinical endpoints. We also compare the effect of inclusion or exclusion of non-tumor tissue regions on prediction performance.

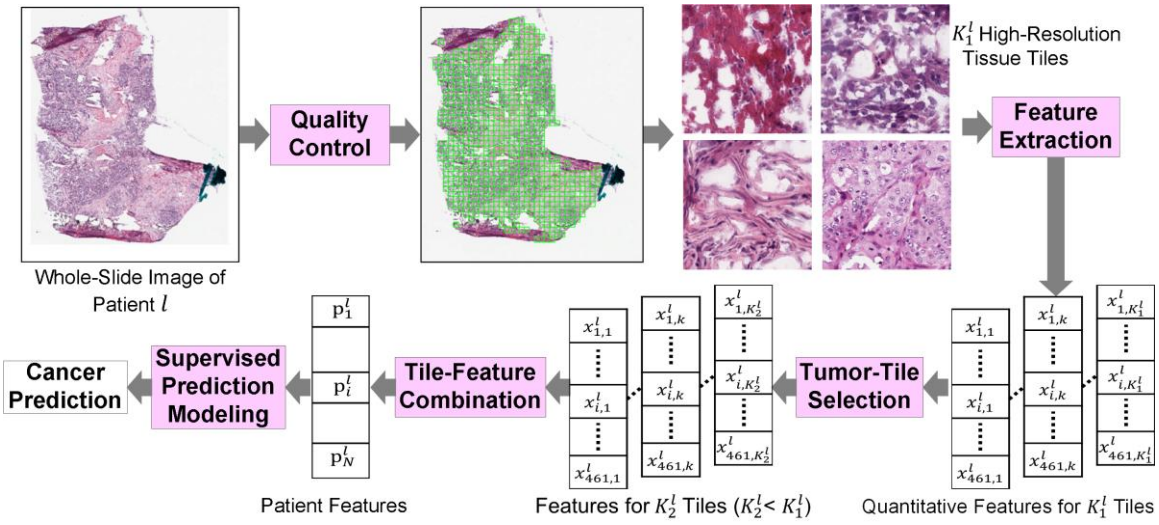

Figure 64: Flow diagram for decision making using whole-slide images.
We select regions in WSIs without tissue folds and pen marks using low resolution images. We then extract 461 image features from high resolution tiles and select tiles from tumor regions based on a supervised model. Thereafter, we represent patients by combining features from tumor tiles using different tile combination methods. Finally, we develop predict models using feature selection and classification.

Materials and Methods

Data

We perform this study using publicly available WSIs of H&E-stained tumor samples of ovarian serous adenocarcinoma (OvCa) and kidney renal clear cell carcinoma (KiCa) from TCGA [3]. TCGA database contains multiple tissue samples for each patient. We use WSIs of 1,092 tumor samples from 563 OvCa patients and 906 tumor samples from 451 KiCa patients. TCGA database also includes several clinical factors (diagnoses) for each patient. Among the reported clinical factors, we use histological grade, metastasis, stage, and survival for KiCa patients and histological grade, survival, stage, and, lymphnode spread for OvCa patients. Based on these clinical factors, we focus on four binary prediction endpoints for both KiCa and OvCa (Table 21). Since clinical

factors are not available for all patients, we use a different number of patients for each clinical endpoint. Out of four different WSI resolutions provided by TCGA, we use the lowest and the highest resolution WSIs for quality control and decision making, respectively.

Table 21: List of clinical binary endpoints of OvCa and KiCa.

Cancer	**Endpoint**	**Class 1**		**Class 2**	
		Description	**Patients**	**Description**	**Patients**
KiCa	Grade	Grade 1 or 2	204	Grade 3 or 4	239
	Metastasis	No spread to other organs	381	Spread to other organs	68
	Stage	Stage I or II	267	Stage III or IV	182
	Survival	< 5 years	126	>=5 years	101
OvCa	Grade	Grade 1 or 2	71	Grade 3 or 4	478
	Stage	Stage I or II	42	Stage III or IV	515
	Survival	< 5 years	252	>=5 years	80
	Lymphnode	No spread to nearby lymph nodes	77	Spread to nearby lymph nodes	134

Quality Control

Using quality control methods described in Chapter 2, we identify blank, pen-mark, and tissue-fold regions in low-resolution WSIs (Figure 65.A). We represent the patient l using a collection of K_1^l high-resolution tissue tiles from quality-controlled regions. We crop an HxW-dimensional WSI into an MxN-dimensional matrix of 512x512-pixel non-overlapping tiles, where $M = \left\lfloor {}^{H}/_{512} \right\rfloor$ and $N = \left\lfloor {}^{W}/_{512} \right\rfloor$. Then, we select tiles with greater than 50% tissue and less than 10% tissue folds. Figure 65.B shows tissue tiles, without artifacts or blank regions, selected for the OvCa WSI in Figure 65.A.

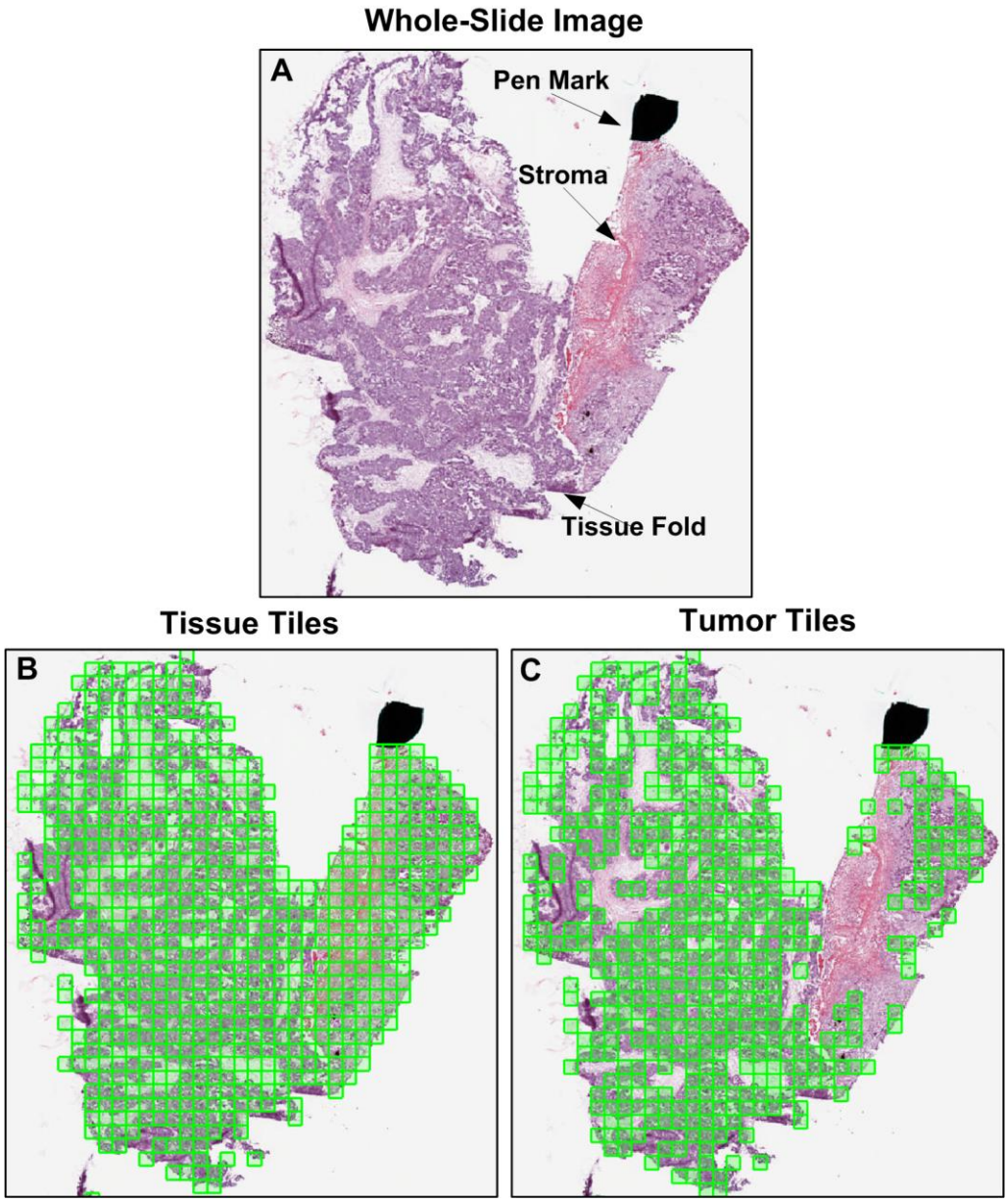

Figure 65: Quality control and tumor selection in a sample OvCa WSI.

Image Feature Extraction

We describe each tissue tile using a comprehensive set of 461 quantitative image features (Table 16) from nine high-level feature subsets: color (C), global texture (GT), basophilic-stained object shape (BOS), eosinphilic-object shape (EOS), no-stain-object shape (NOS), eosinphilic-region texture (ET), basophilic-region texture (BT), nuclear shape (NS), and nuclear topology (NTo). After the feature extraction step, each tile k for patient l using a set of 461 quantitative features is represented by $\mathbf{x}_k^l$.

Tumor Region Selection

We classify tissue tiles into tumor and non-tumor tiles using a supervised model F_A, which is trained on a set of manually annotated tiles. The model maps image features for a tile $\mathbf{x}_k^l$ to an annotation label A_k^l, $A_k^l = \mathrm{F}_A(\mathbf{x}_k^l)$, where $A_k^l = 1$ for tumor and $A_k^l = 0$ for non-tumor tiles. We select the top features using the mRMR-q feature selection method [93] and develop a classification model using the linear-SVM classifier [156]. We incrementally train the model using a framework in which we iteratively added more training tiles (from different WSIs) to provide the classifier with more challenging training samples. We separately optimize F_A for KiCa and OvCa dataset because the morphology of cancer and non-cancer regions differ in the two cancer types. We use 15 and 17 WSIs while training F_A for OvCa and KiCa, respectively. The optimal feature size selected for both models is 17. After tumor region selection, we represent patient l using a collection of K_2^l tumor tiles, where $K_2^l < K_1^l$. Figure 65 shows tumor tiles selected for an OvCa sample. We develop prediction models using both tissue and tumor tiles to

understand the effect of tumor-region selection on the performance of various cancer prediction models.

Tile Feature Combination

In order to classify patients, we need to systematically combine tile features $\mathbf{x}_k^l, k \in [1, K^l]$ to represent each patient, where K^l is equal to K_2^l or K_1^l, indicating inclusion or exclusion of non-tumor regions, respectively. The main objective for a tile combination method is to output patient features $\mathbf{p}^l$ that adequately capture knowledge in the WSI. We compare four tile combination methods: Simple, Univariate Quantization, Multivariate Quantization, and Multivariate Subset Quantization.

Simple Combination (Simple)

For the simple combination (Simple) method, we combine tile features $\mathbf{x}_k^l$ to best approximate the patient features that would have resulted by processing the WSI as a whole (rather than by tiles). Specifically, we consider the collection of tiles as a subset of the WSI and features extracted from tiles are combined using group statistics. The combination method is dependent on the feature type as listed in Table 22. For most features, we directly combine features while for some features (such as fractal), we combine semi-processed data (such as histogram frequencies).

Table 22: Feature-dependent simple combination.

Feature $x_{i,k}^{l}$	**Type of combination**
Count, probability, energy, entropy, and pixel-level-averages for tiles	Average of tiles: $p_i^l = \frac{\sum_k x_{i,k}^l}{\sum_k (1)}$
Object-level maximum	Maximum among tiles: $p_i^l = \{x_{i,k}^l \mid x_{i,k}^l \geq x_{i,l}^l \; \forall \; k \neq l\}$
Object-level minimum	Minimum among tiles: $p_i^l = \{x_{i,k}^l(i) \mid x_{i,k}^l \leq x_{i,l}^l \; \forall \; k \neq l\}$
Object-level average	Group average: $p_i^l = \frac{\sum_k x_{i,k}^l n_k^l}{\sum_k (n_k^l)}$, where n_k^l is number of objects in tile k of patient l
Object-level standard deviation	Standard deviation of a collection of tile objects (group statistics): $p_i^l = \sqrt{\frac{e_i^l + g_i^l}{\sum_k (n_k^l)}}$, where error sum of squares $e_i^l = \sum_k x_{i,k}^l x_{i,k}^l (n_k^l - 1)$ and group sum of squares $g_i^l = \sum_k n_k^l (\mu_{i,k}^l - \mu_i^l)^2$, where $\mu_{i,k}^l$, μ_i^l are tile- and patient-level mean of the object-level image feature, whose standard deviation is being calculated.
Haralick and Fractal features	Combine frequency matrices from each tile k and then calculate features

After simple combination, a patient-level feature may have a numeric value in a range different from other patient-level features. For instance, median nuclear area of all patients in KiCa approximately lies between 200 and 1000 pixels while nuclear eccentricity lies between 0 and 1. This difference in feature range affects feature selection and classification. Therefore, we normalize all features to be in the range [0,1] using rank normalization [163]. We develop a rank function R_i, which orders the feature values $p_i^1, p_i^2, \ldots, p_i^P$ such that $R_i(p_i^m) > R_i(p_i^n)$ implies $p_i^m > p_i^n$. Using this rank function, we normalize the features as follows:

$$\hat{p}_i^l = \frac{R_i\left(p_i^l\right)-1}{P-1}, \tag{50}$$

where P is the number of patients.

Univariate Quantization (UniQ)

For the univariate quantization (UniQ) method, instead of estimating a single value for a feature of a patient, we represent a feature of patient as a histogram with a fixed number of bins (Figure 66). In other words, we first quantize each feature into B bins. For each WSI, we then estimate the percent of its tiles that fall into each bin. Finally, all B percentages for all features represent a single patient profile. Quantization values for a feature are estimated using the number of bins B, which is a parameter, and a feature-dependent dynamic range: lower limit L_i and upper limit U_i . The limits are calculated based on the distribution D_i of the feature i across all tiles of all patients in the training set using quantiles as discussed in equations (47) and (48). In Figure 66, we show the variation in a feature, median nuclear eccentricity, across the tiles of a sample. The color map in the spatial visualization of Figure 66.A corresponds to the quantization bins of the histogram in Figure 66.B.

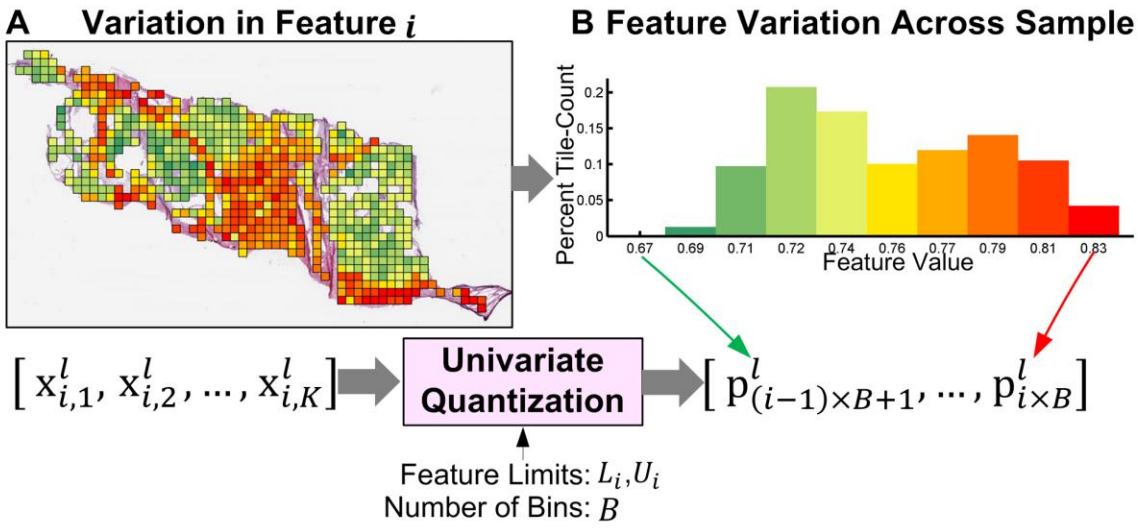

Figure 66: Flow diagram for univariate quantization of an image feature i.
Input parameters for univariate quantization include number of bins and feature limits. The number of bins is pre-selected parameter while feature ranges are calculated using training set. (A) Variation in median nuclear eccentricity across the tiles of KiCa sample. Color of a tile corresponds to its quantization bin. (B) Quantization histogram illustrating percent of tiles in various bins.

Multivariate Quantization (MultiQ)

For the multivariate quantization (MultiQ) method, we quantize a multi-dimensional feature space (instead of a single feature) based on the natural separation between training-set tiles in the feature space. This quantization is non-linear in nature and allows more or less quantization blocks in a subspace of the feature space depending on the feature variation in the subspace. Therefore, unlike UniQ, it can capture knowledge in fewer quantization blocks. MultiQ is very similar to the bag-of-words (BOW) paradigm in computer vision [176], except that, in tile combination, "feature" is a combination of morphological properties of tiles instead of a local key-point descriptor. Moreover, we adopt dense sampling, wherein we combine all tiles, instead of the more commonly used sparse sampling in BOW methods [177].

MultiQ is a four-step procedure. First, we normalize and convert each tile feature into unsigned integers in the range $[0,255]$. For this conversion, we calculate the upper and lower limits (L_i, U_i) for each feature in the training-set tiles in a manner similar to that of univariate quantization and divide the feature range into 256 bins. Thereafter, for each tile, we map the feature value to its bin number. This normalization forces all features to have a similar range [178]. Second, we cluster the training-set tiles into C clusters. These clusters are often referred to as the codebook in the BOW paradigm. A training set can include up to a half-million tiles described by 461 image features. Because of the size of the data, it is not feasible to use simple k-means clustering and we adopt the approximate k-means method for clustering [179]. Approximate k-means uses k-d trees for calculating distances between tiles (nodes) and reduces the computational complexity from $\mathcal{O}(TC)$ to $\mathcal{O}(T \log C)$, where T is number of tiles. We use approximate

k-means with eight randomized k-d trees and five k-means iterations [178]. Third, we map tiles for each patient to the closest cluster among C clusters. Again, we calculate distance using k-d trees [178]. Finally, we calculate the percent of tiles mapped to each cluster. Figure 67 illustrates the multivariate quantization method, where tiles in Figure 67.A are mapped to different clusters as shown in Figure 67.B. We have highlighted bars for two clusters (359 and 455) in the histogram and in corresponding tiles in the WSI. All magenta and cyan tiles are in non-tumor and tumor regions of the WSI, respectively, and correspond to high and low median nuclear eccentricity in Figure 67.A.

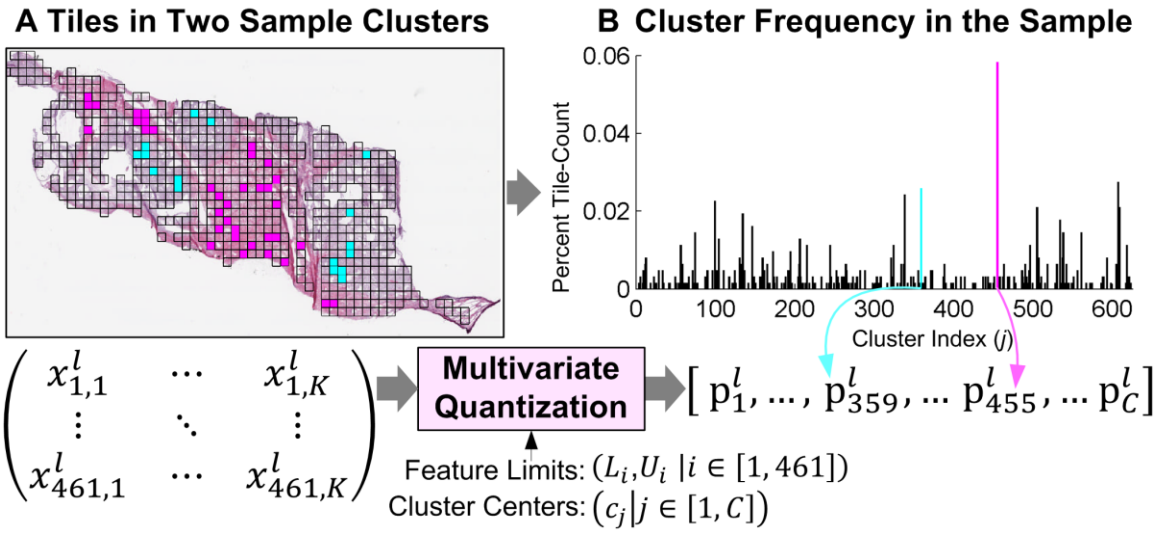

Figure 67: Flow diagram for multivariate quantization of tiles features.
Input parameters for multivariate quantization include feature limits for normalization and pre-computed cluster centers for mapping. Tiles painted magenta and cyan in KiCa sample (A) corresponds to two clusters in the histogram (B). All the tiles in magenta and cyan cluster belong to non-tumor and tumor regions of the sample.

Multivariate Subset Quantization (MultiSubQ)

Since the UniQ method quantizes every feature, it ensures that final patient features include all morphological properties. However, it results in a large number of patient features. On the other hand, MultiQ reduces the number of features but does not ensure that final patient features include all morphological properties. Moreover, it is difficult to biologically interpret patient features resulting from MultiQ because it is difficult to map cluster centers to the original image features. Multivariate subset quantization (MultiSubQ) is a compromise between the two methods. In MultiSubQ, we quantize each of the nine feature subsets into C clusters following the first and second steps of the MultiQ method. Thereafter, we map tile features for each patient to C clusters corresponding to each subset resulting in $9 \times C$ patient features. We found that the required parameter C for MultiSubQ is much smaller than that of MultiQ. Thus, the total number of patient features is much smaller than that of UniQ. Moreover, since we can easily map cluster centers to a high-level feature subset, MultiSubQ is easier to interpret biologically.

Feature Selection and Classification

For each cancer endpoint, we select the most informative patient features and develop prediction models. After tile combination many patient features may be correlated (especially after the UniQ method). Thus, we select features using mRMR feature selection [93]. We develop binary prediction models using classifiers based on discriminant analysis—linear, quadratic, spherical, and diagonal. We optimize feature

size (between 1 to 100) and classifier parameters using 5-fold, 10 iterations of nested CV [180].

Results and Discussion

Impact of Quantization on Prediction Performance

Figure 68.A and Figure 68.B illustrate the performance of prediction models for KiCa and OvCa endpoints. Because of the large difference in prevalence of each class for some endpoints, we use area under the curve (AUC) to measure the prediction performance. Reported AUC is an average performance of predicting testing set samples in 50 external loops of nested CV. The cells in Figure 68.A and Figure 68.B are painted based on their AUC values. Cells with high AUC values are painted red followed by yellow, green, cyan, and blue. The bars on the right and bottom of the figures correspond to average performances for different methods and endpoints, respectively. We report results using three values for each parameter (bin or cluster size) and each quantization method.

For KiCa endpoints, models based on MultiSubQ perform the best, followed by the models based on MultiQ, then UniQ, and then Simple. Models for grade and metastasis prediction perform the best with 0.70 AUC. Models for 5-year survival perform the lowest but the performance improves with quantization. Based on the bars in the bottom of the figure, average performance improves slightly with tumor region selection. The parameters do not have a significant effect on the prediction performance except for MultiSubQ, whose performance improves with increasing number of clusters. For OvCa

endpoints, models based on MultiSubQ features perform best for the grade and survival endpoints while models based on UniQ perform best for the lymphnode and stage endpoints. It is possible that MultiSubQ performance may improve with a larger number of clusters.

Among different OvCa models, the model for the stage endpoint performs best with 0.76 AUC while the model for the grade endpoint performs second best with 0.70 AUC. The performance of models for the survival endpoint is low (close to 0.5, which indicates random prediction) but the performance improves when using MultiSubQ. The performance of MultiQ models slightly decreases with increasing C while the performance of models based on other quantization methods is not considerably affected by parameter values.

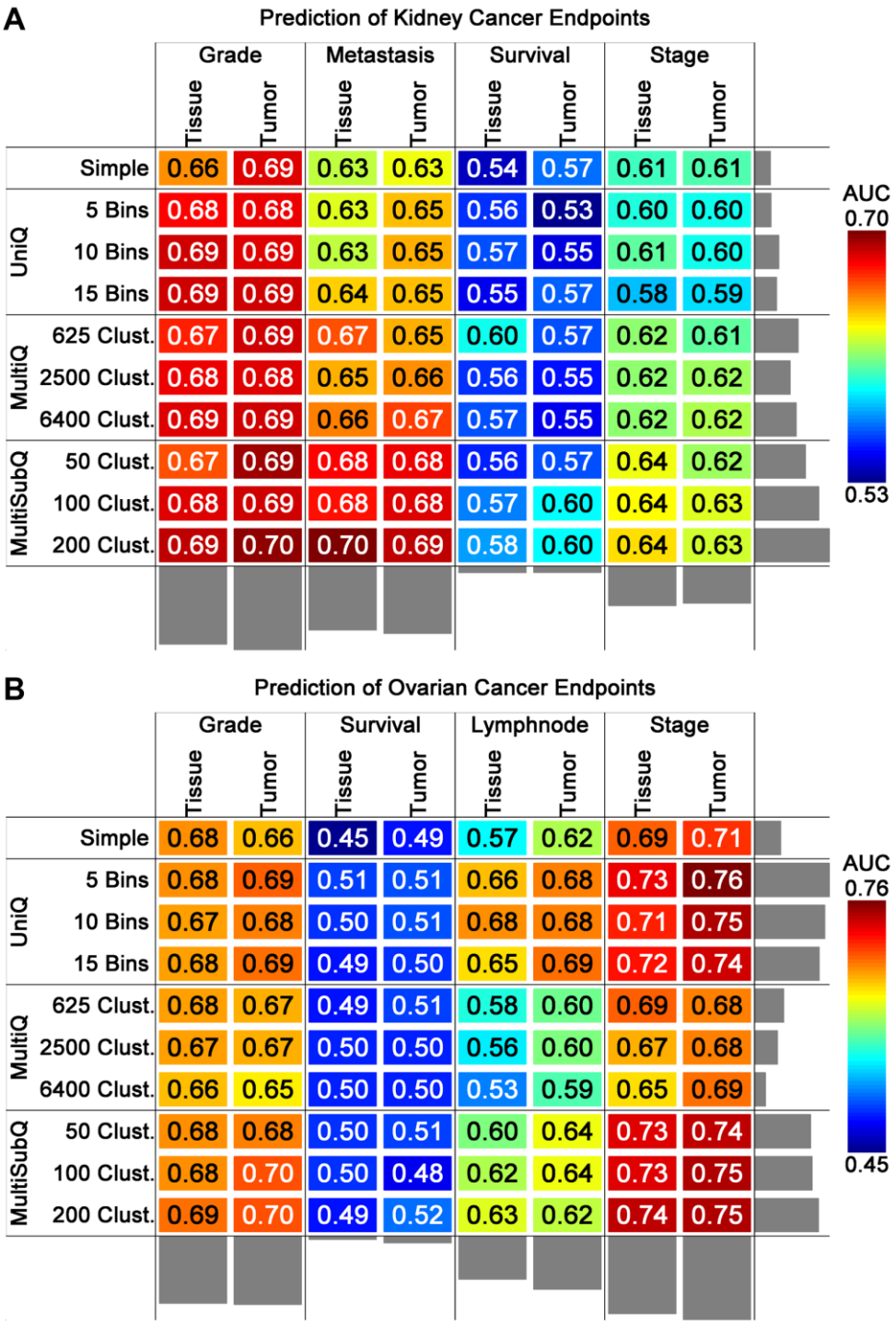

Figure 68: Prediction performance of models based on various types of patient features. For each endpoint, we compare patient features based on tumor and tissue tiles and four tile combination methods. Cells are painted based on AUC values. Bars on the right and bottom of each colormap represent average performance for tile combination methods and endpoints, respectively.

Necessity of Tumor-Region Selection for Diagnosis

In this section, we elaborate on the need for tumor-region selection for WSI-based cancer prediction. Figure 69 shows the change in model performance after tumor-region selection compared to the performance of the model (with the same parameters) without selection. Positive and negative change indicates an increase and decrease in performance after tumor-region selection. Bars in Figure 69 indicate mean change and 99% confidence intervals (CI) of the sample mean, measured across all parameters for 50 iterations of external CV. If the CI of a model does not intersect with the dashed line at zero mean, then the model's performance with and without tumor-region selection is statistically different ($p<0.01$). We can make the following observations: (1) Tumor-region selection improves or maintains performance in most cases with three exceptions: KiCa stage with UniQ, KiCa stage with MultiSubQ, and OvCa grade with Simple; and (2) In most cases, tumor-region selection affects Simple combination more than it does for quantization methods. The decrease in the performance of models with tumor-region selection for the KiCa stage endpoint suggests that non-tumor regions can be informative for KiCa staging. This is interesting because previous work establishes a statistically significant association between tumor necrosis and high stage in kidney renal carcinomas [181]. A large increase in performance for models, especially for KiCa endpoints, based on Simple features compared to quantization methods indicates that quantization methods are more robust in accounting for biological variation in samples by assigning biologically different tiles to different quantization blocks.

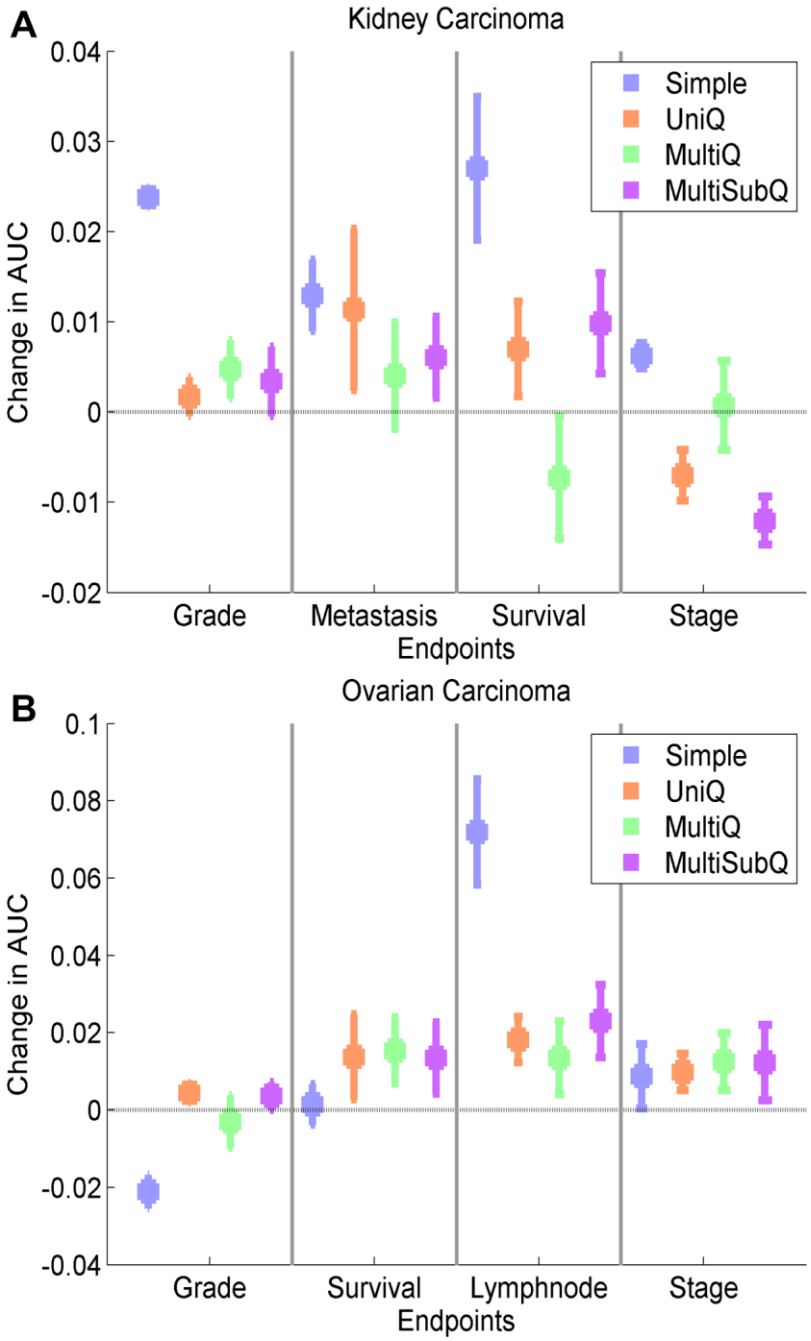

Figure 69: Box plots illustrating change in performance after tumor selection.
Positive and negative change indicates increase and decrease after selection, respectively.

Informative Feature Subsets

Patient features include descriptors from nine high-level subsets (Table 23), among which some subsets are more useful than others for certain endpoints. In this section, we discuss which feature subsets are most informative for various endpoints based on 50 prediction models optimized in nested CV. Optimal models may have different numbers of features (up to 100 features), so we calculate the average percent contribution of each subset across all models. Thereafter, we establish the importance of a feature subset by calculating the p-value for a one-sided Fisher's exact test. Low p-value rejects the null hypothesis that the number of features selected from a subset is equal to random chance. Fisher's exact test is often used to identify over-represented Gene Ontology terms in a list of genes [135]. Table 23 reports p-values for all feature subsets in the best performing models for each endpoint. We have highlighted statistically over-represented feature subsets for each endpoint ($p<0.05$, adjusted for multiplicity using the Bonferroni method).

Table 23: Statistically over-represented image features subsets in OvCa and KiCa clinical diagnosis models using WSIs.

Subset	**Grade**		**Stage**		**Survival**		**Metastasis**	**Lymphnode**
	KiCa	**OvCa**	**KiCa**	**OvCa**	**KiCa**	**OvCa**	**KiCa**	**OvCa**
C	0.074	0.023	0.000	0.415	0.000	0.052	0.000	0.399
GT	0.550	0.565	0.942	0.985	0.691	0.007	0.202	0.195
EOS	0.990	0.691	0.942	0.309	0.991	0.341	0.973	0.970
ET	0.550	0.135	0.565	0.555	0.691	0.716	0.202	0.545
NOS	0.938	0.002	0.314	0.011	0.885	0.820	0.678	0.668
BOS	0.879	0.801	0.942	0.309	0.997	0.979	0.997	0.298
BT	0.021	0.942	0.135	0.760	0.000	0.092	0.301	0.344
NS	0.001	0.991	0.565	0.828	0.975	0.979	0.791	0.924
NTo	1.000	0.997	0.975	0.785	0.997	0.992	0.990	0.634

Since pathologists grade KiCa based on the Fuhrman nuclear grading system, over-representation of nuclear shape features confirms that our models are based on image properties that are also used by pathologists for manual analysis [182]. Similarly for OvCa, with progression in cancer grade, cells become less differentiated and tissues lose their serous property (i.e., a property in which cavities fill with serous fluid). Thus, our prediction models include no-stain object shape features, which capture serous structures. Traditionally, pathologists diagnose metastasis, stage, and lymphnode spread based on gross analysis of tumors instead of by histopathological analysis. In our study, we found that metastasis and stage can be predicted with reasonable accuracy based on histopathological properties. Color properties were most informative for KiCa metastasis and stage while basophilic-object shape properties were most informative for OvCa stage. For KiCa survival, we found that color and basophilic texture were statistically over-represented.

Effect of Tissue-Fold Artifact on Cancer Grading

Previous studies have discussed the need for eliminating tissue-fold image artifacts before extracting image features and building diagnostic models. However, to the best of our knowledge, no published work investigates the effect of tissue folds on quantitative image features and cancer diagnosis. We identify image features changed by the presence of tissue folds using a rank-sum test of two lists of feature values in WSIs with and without tissue folds. If the p-value for the test is less than 2.1692e-005 (i.e., a p-value threshold of 0.01, adjusted for multiplicity by the Bonferroni method, 0.01/461), then this indicates that a feature changes (with statistical significance) in the presence of tissue

folds. Figure 70 shows several image features changed by tissue folds. We found that 30 and 53 features changed by tissue folds in OvCa and KiCa, respectively. Out of these features, most capture an extreme value or variation in a property such as the minimum averaged distance of a nucleus to its five neighbors and the standard deviation in a nuclear area. Moreover, the presence of tissue folds increases the spread of most features. Hence, tissue-fold artifacts create outlier regions in a WSI.

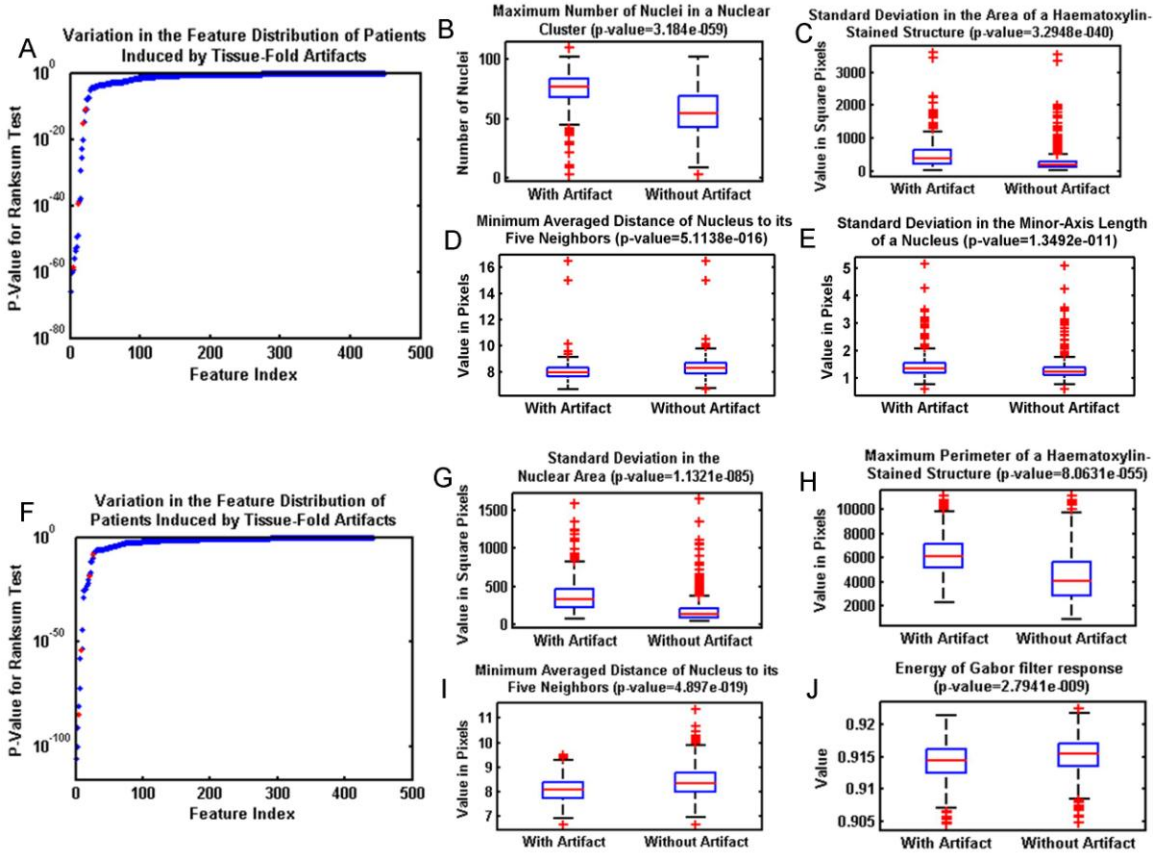

Figure 70: Effect of tissue-fold elimination on quantitative image features. Variation in quantitative image features in the WSIs of KiCa (A-E) and OvCa (F-K) samples with the presence of tissue folds. The p-value of the rank-sum test of lists of feature values with and without folds (A & F). With the presence of tissue folds, 30 and 53 image features statistically changed in KiCa and OvCa, respectively. Using box-plots, we illustrate the distribution of certain features (highlighted in red) changed by tissue folds.

To investigate the effect of tissue folds on predictive grading models, we develop KiCa and OvCa predictive grading models for WSIs using the features changed by tissue folds. We found that, without folds, these models have higher AUC, as assessed with the ten iterations of 5-fold of CV (Table 24).

Table 24: AUC of predictive grading models with and without tissue folds.

Cancer	Without Fold	With Fold
OvCa	0.59±0.01	0.54±0.03
KiCa	0.66±0.01	0.65±0.01

The improvement in AUC is more prominent for the OvCa data set, which includes WSIs with a higher percent of tissue folds (Figure 71). Therefore, we can conclude that the presence of tissue folds changes several quantitative image features. Consequently, after eliminating the tissue-fold regions in WSIs, prediction models based on these image features can more accurately classify WSIs into groups of high and low grade.

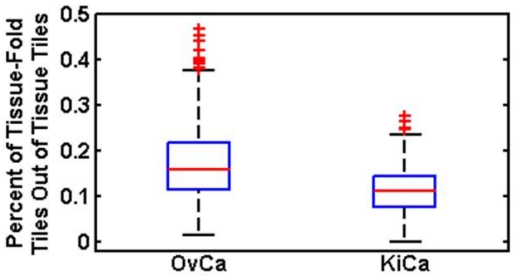

Figure 71: The percentage of tissue folds in WSIs provided by TCGA.
The value on the y-axis represents the percentage of tissue tiles eliminated because of tissue folds in samples per patient. Samples for patients with OvCa have more tissue folds than patients with KiCa.

Conclusion

An important challenge for pathology imaging informatics is to capture the knowledge in large histopathological whole-slide images (WSIs) which can aid in decision making. A simple method for WSI representation averages image features across a complete image. In this chapter, we proposed and tested three novel WSI representation methods that quantify the percent of different biological regions in a WSI using feature space quantization. We compared these methods using four kidney and four ovarian carcinoma endpoints from publicly available datasets. Our results indicate that quantization-based features improve decision making by up to 7% AUC. We also compared the performance of different methods when features were extracted from only the tumor regions of a WSI. We found that quantization methods are less sensitive to tumor-region selection compared to the simple method. Moreover, tumor-region selection reduces the performance of kidney cancer staging. We found that statistically over-represented feature subsets for grading endpoints correspond to image properties that a pathologist would normally identify in a traditional diagnosis. We found that presence of tissue-fold artifacts change simple image features and decrease OvCa and KiCa grading performance. Our methods can be easily extended to other cancer endpoints and represent an important step towards effective and automated WSI-based clinical decision-making.

CHAPTER 10

PATIENTVIZ: VISUALIZATION TOOL FOR PATIENT STRATIFICATION

Introduction

This chapter addresses the informatics challenge of optimizing and validating prognosis prediction models. It discusses the development of an interactive patient-level visualization tool, PatientViz, which allows user to study patient stratification in terms of prognostic significance, stability, and reproducibility simultaneously. The chapter also develops a method for genomic stratification using histopathological knowledge.

Cancer is one of the leading causes of death in United States. Because of early detection and targeted treatment, cancer mortality rate has decreased 20% from 1991 to 2009 [183]. To further decrease the mortality rate, it is essential to identify novel biomarkers for sub-classification of the disease and develop drugs for targeted treatment of different disease sub-groups. Researchers have suggested machine learning methods for discovering biomarkers using high-throughput genomic data [184-187]. These methods can be broadly classified into supervised and unsupervised methods. A supervised method segregates patients into groups based on their diagnosis and selects a subset of most informative genes. In contrast, an unsupervised method segregates patients into groups based on the dissimilarity in patient's genomic profile and selects a subset of most informative genes. Supervised techniques are useful for clinical prediction modeling while unsupervised techniques are more appropriate for exploratory analysis [188]. Unlike supervised methods, unsupervised methods may result into a grouping that

is not biologically useful [189]. Moreover, unsupervised patterns are often unstable and overfitted to the data [190]. Therefore, genomic clusters are often unstable with limited number of samples [189].

Recently, development of computational methods for biomarker discovery has become an active area of research because of the high-throughput genomic, proteomic, and imaging data acquisition technologies. Several large-scale datasets are publicly available for research at repositories such as Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). TCGA, a joint project by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), provides high-quality genomic, proteomic, and imaging data for same patients [3]. TCGA includes data for hundreds of patients from 20 different cancer types. Using TCGA data, researchers discovered genomic and imaging biomarkers for different cancer types. Researchers discovered prognostically significant stratification of glioblastoma patients using genomic [184, 185, 187] and imaging datasets [65, 110]. Liu et al. associated gene expression and nuclear image profiles with chemotherapy response in OvCa [168]. Fackler et al identified biomarkers specific to breast cancer hormone receptor status and recurrence risk using genome-wide methylation analysis [186].

A patient stratification should be stable and reproducible for it to be clinically useful. The task of assessing the quality of clustering is non-trivial because even random data can result in a stable (but irreproducible) clustering [190]. Researchers have proposed several heuristic and mathematical metrics for measuring the quality of

clustering [190, 191]. However, none of these metrics is widely accepted and researchers often study multiple metrics in combination to find optimal clustering [192].

We propose a novel algorithm for genomic stratification which is a combination of supervised and unsupervised methods. Using unsupervised clustering of histopathological features, we segregate patients into groups. Thereafter, we use the group labels and genomic data to train a supervised stratification model. Our proposed method assumes that in a stable patient stratification, patients should differ on both histopathology and genomic levels. Using genomic and imaging data provided by TCGA, we develop patient stratification models for ovarian and renal carcinoma. We also develop a graphical interface, called PatientViz, to study various global-stability and cluster-reproducibility metrics simultaneously.

Methods and Materials

Datasets

We use genomic RNA-Seq and whole-slide imaging data of OvCa and KiCa patients provided by TCGA. For stratification and model optimization, we only use patients that have both imaging and genomic data. We use remaining genomic and imaging samples as references during normalization and feature extraction. Table 25summarizes number patients used in main study and normalization for both cancers. We represent WSIs using univariate quantization frequency features with 10 quantization bins and tumor selection, as described in Chapter 9. We normalize RNA-Seq data using trimmed mean of M values (TMM) normalization [193]. To avoid any batch effects, we

divide main-study samples into train and test sets using stratified sampling based on acquisition site.

Table 25: Number of patients in different datasets.

Dataset	**KiCa**	**OvCa**
Histopathological Reference	27	307
Genomic Reference	45	6
Histopathological/ Genomic Train	282	171
Histopathological/ Genomic Test	140	85

Unsupervised Histopathological Clustering

We cluster patients in train set using histopathological features. We use a two-step clustering procedure: (1) feature selection and (2) clustering.

Feature Selection

Feature selection methods for unsupervised learning maximize information in a subset of features irrespective of their relevance to the biological problem and then look for inherent patterns [194]. We use a method based on SVD-entropy to select the most informative subset among 4610 univariate-quantization features [194]. We adopt the following two leave-on-out feature selection methods: (1) Simple, and (2) Forward [194]. During simple selection, we calculate contribution of each feature i to the entropy using

$$CE_i = E(A_{N\times 4610}) - E\left(A^i_{N\times 4609}\right), \quad (51)$$

$$E\left(B_{N\times J}\right) = \sum_{k=1}^{N} V_k \log V_k, \quad (52)$$

where A is a matrix of all features for N samples, A^i is a matrix of all features except feature i and $V_k = s_k^2/\sum_l s_l^2$, s_l^2 are the eigenvalues of the NxN matrix BB^T. Thereafter, we rank features based on decreasing CE_i. High values of CE_i signifies that the presence of feature i adds useful dimension to the data and distributes information over multiple Eigen vectors [194]. During forward selection, we follow following steps: (1) select a feature i in A with maximum CE_i (2) remove feature i from A matrix, and (3) follow steps (1) and (2) till required number of features is selected.

Clustering

We use agglomerative hierarchical clustering to cluster patients in a multidimensional image-feature space. Based on feature ranks, we select top 400 image features and cluster patients using 16 different feature sizes including 25, 50, …, and 400. We cluster similar patients using two criteria: (1) Ward linkage with Euclidian distance and (2) Average linkage with correlation distance. Ward's linkage minimizes the increase of the within-cluster sum of squares as a result of merging two clusters [171]. The increase in sum of squares by merging clusters *k* and *l* is measured by the following distance metric:

$$D_w(k,l) = \sqrt{\frac{2n_k n_l}{n_k+n_l}}\|c_k - c_l\|, \tag{53}$$

$$c_k = \frac{1}{n_k}\sum_{r=1}^{n_k} x_{kr}, \tag{54}$$

where n_k is the number of patients in cluster *k*, c_k is the centroid of cluster *k*, x_{kr} is m-dimensional representation of patient r in cluster k and $\|.\|_2$ is the Euclidian distance. Average linkage measures distance between two clusters *k* and *l* using average pairwise distance between all objects in the clusters given by

$$D_a(k,l) = \frac{1}{n_k n_l}\sum_{r=1}^{n_k}\sum_{t=1}^{n_l} d_c(x_{kr}, x_{lt}), \quad (55)$$

$$d_c(x_{kr}, x_{lt}) = 1 - \frac{(x_{kr}-\bar{x}_{kr})(x_{lt}-\bar{x}_{lt})'}{\sqrt{(x_{kr}-\bar{x}_{kr})(x_{kr}-\bar{x}_{kr})'}\sqrt{(x_{lt}-\bar{x}_{lt})(x_{lt}-\bar{x}_{lt})'}}, \quad (56)$$

$$\bar{x}_{kr} = \frac{1}{m}\sum_{j=1}^{m} x_{kr}^{j}, \quad (57)$$

where we measure distance between two m- dimensional objects x_{kr} and x_{lt} using correlation. We cluster patients using nine cluster sizes including 2, 3, …, 10. We generate clusters with all possible combinations of following factors (1) Histopathological features with and without tumor selection, (2) simple and forward feature selection methods, (3) 16 feature sizes, (4) nine cluster sizes, and (5) Ward and average clustering linkages. Thereafter, we select optimal clustering using cluster-assessment metrics.

Patient Cluster Assessment

Cluster Quality

We measure cluster quality by assessing clustering patterns irrespective of biological knowledge. Researchers have proposed several metrics to assess cluster quality, which measure correspondence between clustering and true structure in the data [191]. In our study, we measure two types of cluster-quality metrics: validity and reproducibility. Validity measures intra-cluster cohesion and inter-cluster separation while reproducibility measures predictive strength of the clustering pattern.

We use the following measures of cluster validity:

(1) Global silhouette width (GS): It measures how well a patient lies within its clusters. To measure GS, we first measure silhouette width s_{kr} for each patient r in

cluster C_k and silhouette width S_k for each cluster C_k [191]. For a given cluster C_k (k=1,2,…,C):

$$\text{GS}=\frac{1}{C}\sum_{k=1}^{C} S_k, \tag{58}$$

$$S_k = \frac{1}{n_k}\sum_{r=1}^{n_k} s_{kr}, \tag{59}$$

$$s_{kr} = \frac{b_{kr}-a_{kr}}{\max(b_{kr},a_{kr})}, \tag{60}$$

$$a_{kr} = \frac{1}{n_k-1}\sum_{i\neq r, i\in[1,n_k]} d(x_{kr}, x_{ki}), \tag{61}$$

$$b_{kr} = \min_{i\neq r, i\in[1,n_k]} d(x_{kr}, x_{ki}), \tag{62}$$

where x_{kr} is representation of patient r in cluster k; and $d(\cdot)$ is Euclidean and correlation distance with Ward and Average linkages, respectively. Larger global silhouette width (GS) value indicates the better clustering.

(2) Dunn's index (D): It is a function of inter-cluster distances and intra-cluster distances given by

$$D = \min_{i\in[1,C]}\left\{\min_{j\in[1,C], i\neq j}\left(\frac{d(c_j,c_i)}{\max_{k\in[1,C]}\left\{\frac{1}{n_k}\sum_{r=1}^{n_k} d(c_k,x_{kr})\right\}}\right)\right\}, \tag{63}$$

where $c_k$is the centroid of cluster k with n_k patients. Larger D value indicates the better clustering [191].

(3) Davies-Bouldin index (DB): It measures dispersion of patients within a cluster relative to inter-cluster distances [191]. Lowest DB value indicates the best clustering. Davies-Bouldin index is given by

$$DB = \frac{1}{C}\sum_{k=1}^{C} \max_{i\in[1,C], i\neq k}\left(\frac{\frac{1}{n_k}\sum_{r=1}^{n_k} d(c_k,x_{kr})+\frac{1}{n_i}\sum_{r=1}^{n_i} d(c_i,x_{ir})}{d(c_k,c_i)}\right). \tag{64}$$

We measure the reproducibility of a clustering by comparing similarity of clusterings generated using a subset of samples in train set. To calculate reproducibility indices, we follow the following steps: (1) Randomly divide patients in train set into two equal size groups, (2) Cluster both groups into C clusters, (3) map patients in first group to the cluster centers of second group and vice versa, (4) Compare the similarity of original clustering labels of first (second) group to the labels assigned using second (first) group's cluster centers, and (5) repeat steps (1) to (4) 50 times and average similarity indices. We measure the similarity of clusterings using the following indices:

(1) Rand index (RI): It counts the number of agreements in patient pairs between two clusterings. For example, a patient pair in which both patients are clustered as being in the same group in one clustering agrees with the other clustering if both the patients are closest to same cluster in other clustering (clusters determined using remaining half of patients). Alternatively, if patients in the pair are closest to different clusters in the other clustering, the pair is in disagreement. The Rand index is the ratio of the number of agreeing pairs to the total number of pairs given by [195]:

$$RI = A/\binom{N}{2}, \tag{65}$$

$$A = \binom{N}{2} + 2\sum_{k=1}^{C}\sum_{l=1}^{C}\binom{N_{kl}}{2} - \left(\sum_{k=1}^{C}\binom{N_{ko}}{2} + \sum_{l=1}^{C}\binom{N_{ol}}{2}\right), \tag{66}$$

where N_{kl} is number of patients that are in cluster k of first clustering and cluster l of second clustering; N_{ko} and N_{ol} is number of patients in cluster k in first clustering

and cluster l in second clustering, respectively; N is the total number of patients. Higher value of Rand index indicates better clustering.

(2) Mirkin's index (MI): It counts the number of disagreements in patient pairs between two clusterings. It is the ratio of the number of disagreeing pairs to the total number of pairs [195] given by

$$MI = D/\binom{N}{2} \text{ and} \tag{67}$$

$$D = \sum_{k=1}^{C}\binom{N_{ko}}{2} + \sum_{l=1}^{C}\binom{N_{ol}}{2} - 2\sum_{k=1}^{C}\sum_{l=1}^{C}\binom{N_{kl}}{2} \tag{68}$$

Lower value of Mirkin's index indicates better clustering.

(3) Hubert's index (HI): It calculates the difference between the number of agreements and disagreements in patient pairs between two clusterings. It is a combination of Rand and Mirkin's index [195].

$$HI = (A - D)/\binom{N}{2} \tag{69}$$

(4) Adjusted rand index (ARI): Adjusted Rand index is a modification of the Rand index to account for expected value of index depending on group prevalence [195].

$$\text{ARI} = \frac{\text{RI} - \text{E(RI)}}{\max(\text{RI}) - \text{E(RI)}} \tag{70}$$

$$\text{E(RI)} = 1 + 2\sum_{l=1}^{C}\binom{N_{ol}}{2}\sum_{k=1}^{C}\binom{N_{ko}}{2}/\binom{N}{2}^{2} - \left(\sum_{k=1}^{C}\binom{N_{ko}}{2} + \sum_{l=1}^{C}\binom{N_{ol}}{2}\right)/\binom{N}{2} \tag{71}$$

$$max(RI) = 1 \tag{72}$$

$$ARI = \frac{\sum_{k=1}^{C}\sum_{l=1}^{C}\binom{N_{kl}}{2} - \sum_{l=1}^{C}\binom{N_{ol}}{2}\sum_{k=1}^{C}\binom{N_{ko}}{2}/\binom{N}{2}}{\frac{1}{2}\left(\sum_{k=1}^{C}\binom{N_{ko}}{2} + \sum_{l=1}^{C}\binom{N_{ol}}{2}\right) - \sum_{l=1}^{C}\binom{N_{ol}}{2}\sum_{k=1}^{C}\binom{N_{ko}}{2}/\binom{N}{2}} \tag{73}$$

Prognostic Significance

A high-quality clustering pattern may not be biologically useful. We measure the biological (prognostic) significance of a clustering using two-sided log-rank test on pairs of survival functions for patient groups (clusters). Let $t = 1, \cdots, T$ be unique time points when a death occurs in either good- or bad-survival groups; $d_{G,t}$ and $d_{B,t}$ be number of deaths in good- and bad-survival groups, respectively; $a_{G,t}$ and $a_{B,t}$ be the number of patients alive (non-censored) in the two groups; and $d_t = d_{G,t} + d_{B,t}$ and $a_t = a_{G,t} + a_{B,t}$ be total number of patients dead and alive, respectively. The expected number of deaths for bad survival group $E_{B,t}$ and variance V_t under the null hypothesis that both groups have similar survival functions is given by hypergeometric distribution as follows:

$$E_{B,t} = \frac{a_{B,t}}{a_t} d_t \tag{74}$$

$$V_t = \frac{d_t a_{B,t} a_{G,t} (a_t - d_t)}{(a_t - 1) a_t a_t} \tag{75}$$

The log-rank statistic for bad survival z_B compares $d_{B,t}$ to its expected value $E_{B,t}$ under the null hypothesis, mathematically given by

$$z_B = \frac{(\sum_t d_{B,t} - E_{B,t}) - 0.5}{\sqrt{\sum_t V_t}}. \tag{76}$$

The factor of 0.5 is for Yate's correction for continuity. We calculate the statistical significance for z_B using normalized cumulative probability density of a standard normal distribution F, given by $p_1 = F(-z_B)$. Since bad and good survival patients groups are

not known based on clustering, we use two-tailed test to calculate statistical significance as follows:

$$p_2 = 2 \times F(-z), \tag{77}$$

$$z = \frac{|\sum_t d_{B,t} - E_{B,t}| - 0.5}{\sqrt{\sum_t V_t}}, \tag{78}$$

where either of the two groups is selected as the bad-survival group.

Genomic Prediction Modeling

We develop genomic prediction models based on train patients using RNA-Seq genomic as data and clustering groups as label. We develop the prediction models for prognostically significant (two-sided logrank test p-value >0.05) histopathological clusterings with only two clusters: good and bad survival patients. Figure 72 illustrates a block diagram for our genomic modeling method using histopathological clustering. We use mRMR feature selection method [93]. We develop binary prediction models using classifiers based on discriminant analysis—linear, quadratic, spherical, and diagonal. We optimize feature size (range: 1 to 100) and classifier parameters using 5-fold, 10 iterations of nested CV. We also develop a genomic survival-prediction model using five-year survival labels instead of clustering labels. The performance of five-year survival model serves as a baseline for result interpretation.

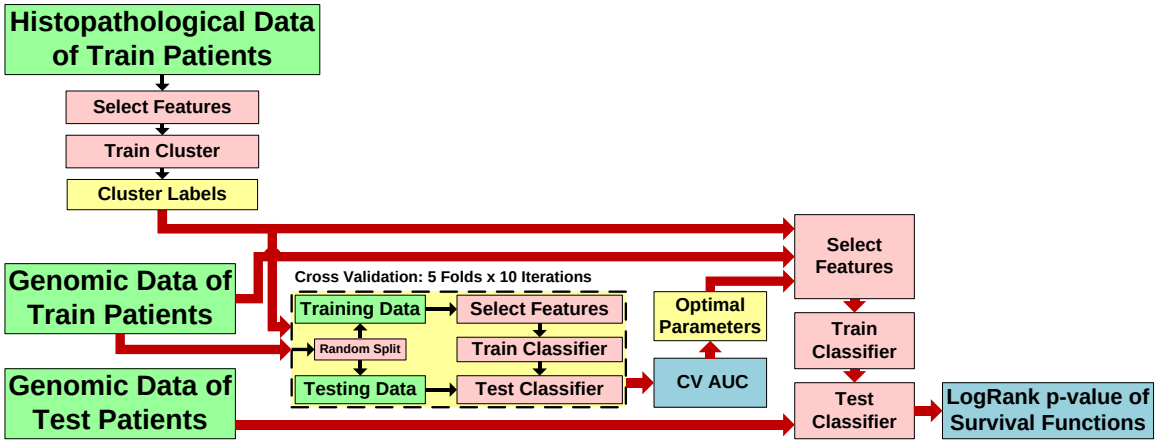

Figure 72: Block diagram for genomic prediction modeling using histopathological knowledge.
Patients in train set are clustered into two groups using histopathological features. Thereafter, a genomic prediction model is trained and optimized using genomic data and clustering group labels. The genomic prediction model is finally tested on a test set using one-sided log rank test on survival functions.

Graphical Tool Design

We develop an interactive visualization tool PatientViz to study patient clustering patterns using histopathological features (Figure 73). PatientViz allows users to simultaneously study three type metrics of clusterings: prognostic significance (Figure 73.A), validity (Figure 73.B), and reproducibility (Figure 73.C). In each heatmap (Figure 73.A-C), feature size and cluster size varies along x-axis and y-axis, respectively. These heatmaps are color coded such that dark red is the best clustering while dark blue is the worst clustering User can select feature and cluster sizes by simply clicking on a box on either of the three heatmaps. For a selected clustering (highlighted in magenta), we can study Kaplan Meier curves (Figure 73.D), similarity of clustering labels to clinical factors (Figure 73.F), and clustering heatmap (Figure 73.G). The clustering heatmap (Figure 73.G) have patients and features along x-axis and y-axis, respectively. Each feature (row)

in the heatmap is standardized such that mean feature value is zero and standard deviation is one. Cells in the heatmap are colored such that value above and below the mean are red and green respectively.

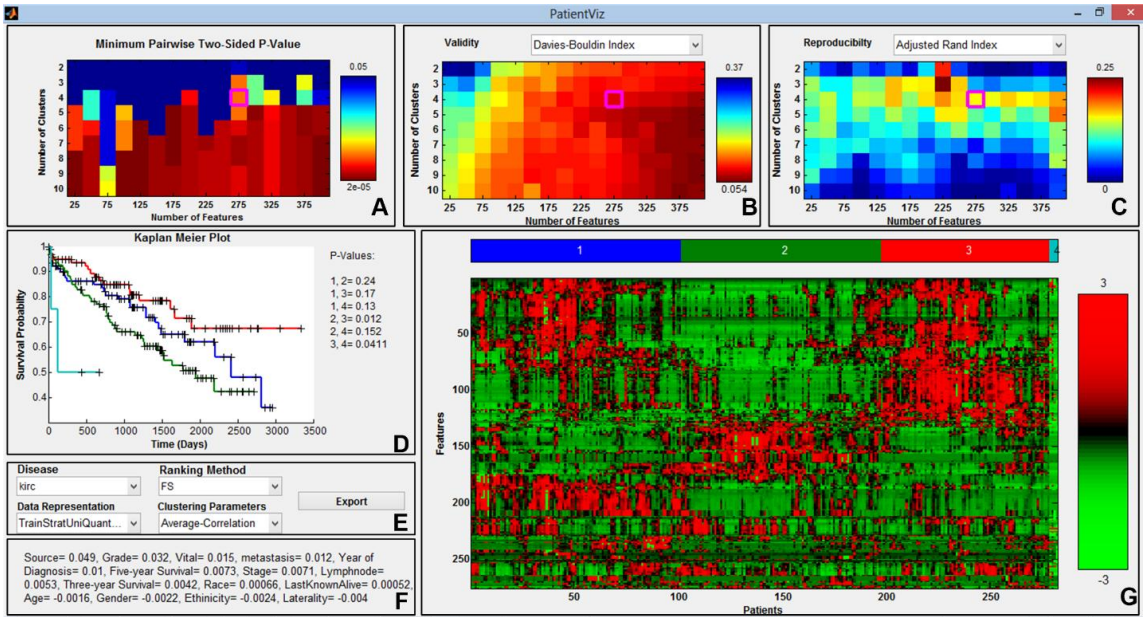

Figure 73: PatientViz- an interactive tool to investigate cancer patient stratification using histopathological features.
(A) prognostic significance, (B) validity, (C) reproducibility, (D) Kaplan Meier curves, (E) feature selection and clustering methods, (F) similarity (adjusted rand index) to clinical factors, and (G) clustering heatmap.

Results and Discussion

We discuss patient stratification results for KiCa and OvCa.

Kidney Cancer Patient Stratification

Survival Prediction Modeling using Five-year Survival Information

In this section, we discuss the performance of the genomic survival-prediction model for KiCa patients trained using RNA-Seq data and clinical five-year survival information. Among train patients (Table 25), 79 and 66 patients have less than and more than (or equal to) five-year survival, respectively. We found that genomic data was informative for survival prediction and model performed with average 0.66 AUC during CV on train set (Figure 72). When we applied the optimized model to stratify patients in the test set, we found that the survival functions for good and bad survival groups are not statistically significant i.e. p-value for one-sided logrank test is greater than 0.05 (Figure 74).

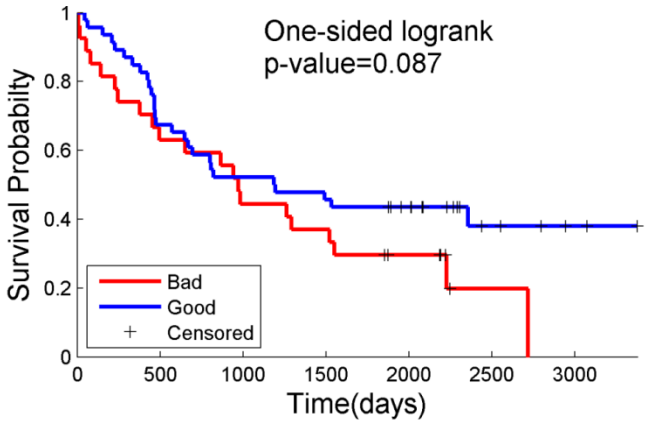

Figure 74: Survival functions of KiCa (test) patients based on a genomic prediction model trained using five-year survival information of train patients.
Survival functions are separated to some extent but not statistical significant (one-sided p-value of logrank test > 0.05).

Association between Cluster Assessment Metrics and External-Validation Performance

Using grouping labels of all prognostically significant histopathological clusterings, we train genomic prediction models. We test these models and calculate one-sided logrank p-value of survival functions. Table 26 lists the correlation of the p-value of test survival functions to all cluster-assessment metrics as well as CV AUC for genomic prediction models. Metrics should be positively or negatively correlated to the p-value given that lowest or highest value of the metric is linked with the best clustering.

We found that most metrics do not measure biological usefulness of a clustering and they are correlated in opposite direction than expected. As expected, genomic CV AUC is negatively correlated to the test p-value and logrank p-value of histopathological clustering is positively correlated to the test p-value. Figure 75 illustrates the relationship of these two metrics with the test p-value. These scatter plots have 21 points with logrank p-value of histopathological clustering less than 0.1 and genomic CV AUC in range 0.69 to 0.80. We select clustering with prognostically significant grouping (i.e. logrank p-value of histopathological clustering <0.05) and maximum genomic CV AUC as the optimal choice for validation. The optimal clustering has following parameters: histopathological features with tumor selection, forward feature selection, 250 feature size, and ward linkage.

Table 26: Correlation between cluster assessment metrics and testing performance.

Metrics based on Train Set	Expected Direction of Correlation	Logrank p-value of survival functions based on Genomic Prediction of test patients
Genomic CV AUC	Negative	-0.248 (p=0.278)
Logrank p-value of Histopathology Clustering	Positive	0.486 (p=0.025)
Adjusted Rand Index	Negative	0.338 (p=0.134)
Rand Index	Negative	0.331 (p=0.142)
Hubert Index	Negative	0.331 (p=0.142)
Mirkin Index	Positive	-0.331 (p=0.142)
Global Silhouette Width	Negative	0.203 (p=0.378)
Davies-Bouldin Index	Positive	-0.175 (p=0.448)
Dunn's Index	Negative	0.062 (p=0.789)

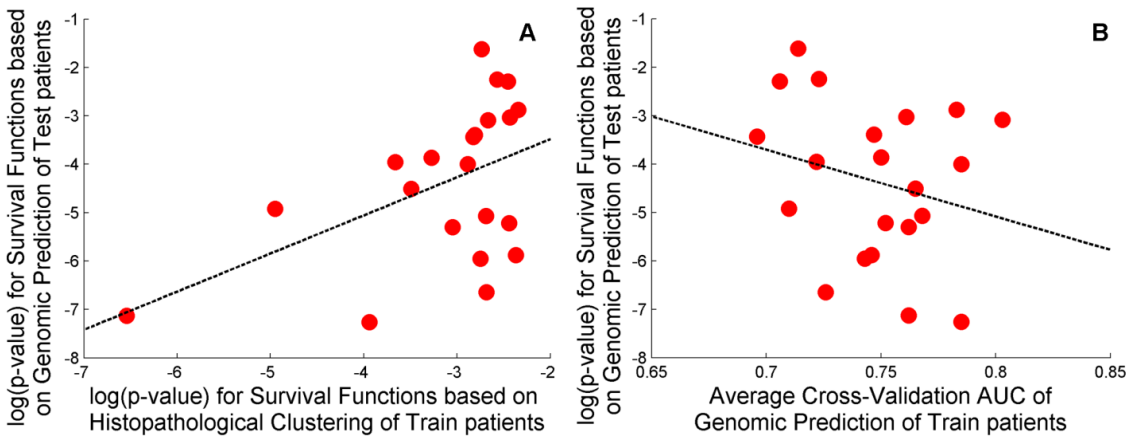

Figure 75: Relationship between cluster assessment metrics and genomic model prediction performance.
(A) Scatter plot of two-sided logrank p-value for survival functions based on histopathological clustering of train patients (x-axis) and one-sided logrank p-value for survival functions based on genomic prediction of test patients (y-axis). (B) Scatter plot of CV AUC of genomic prediction of train patients (x-axis) and one-sided logrank p-value for survival functions based on genomic prediction of test patients (y-axis).

Validation of Optimal Patient Stratification and Informative Biomarkers

We validate optimal histopathological clustering using a genomic prediction model and report informative genomic and histopathological characteristics of this stratification. Figure 76 illustrates Kaplan Meier survival functions of patients in (A) train set based on histopathological clustering, (B) test set based on histopathological prediction, and (C) test set based on genomic prediction. Separation between survival functions based on histopathological clustering and genomic prediction are statistically significant ($p<0.05$). However, separation between survival functions based on histopathological prediction was not statistically significant. Possible reasons for the low performance of histopathological prediction are model overfitting and batch effects.

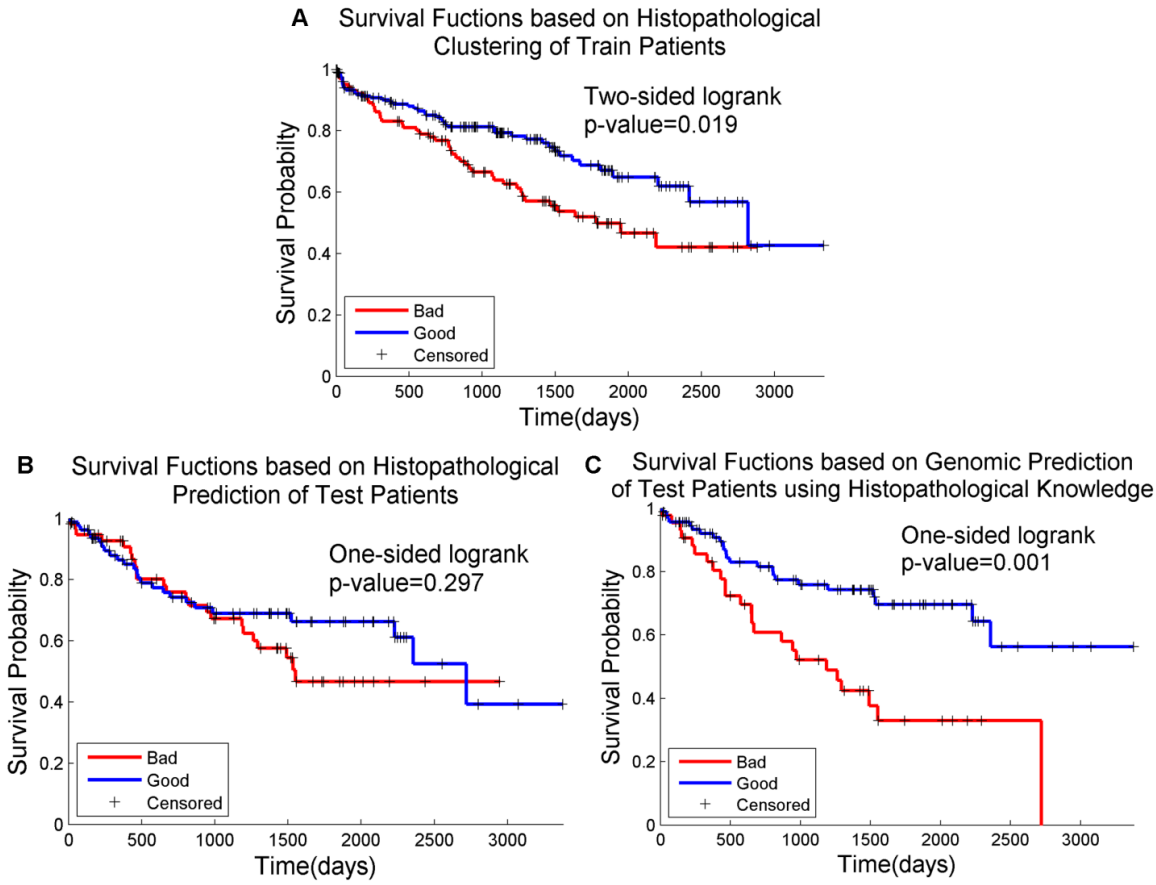

Figure 76: Kaplan Meier curves for KiCa patients stratified using histopathological and genomic properties.
(A) Survival functions for train-set patients stratified into good and bad survival groups using histopathological clustering. (B) Survival functions for test-set patients stratified into good and bad survival groups using histopathological prediction model based on clustering labels. (C) Survival functions for test-set patients stratified into good and bad survival groups using genomic prediction model based on clustering labels. Patient stratifications in train set using histopathological clustering and test set using genomic prediction are statistically significant.

Pathologists often predict survival using clinical factors such as grade, stage, metastasis, and lymphnode spread. A multivariate model depending on these clinical factors have been previously used to predict survival of renal clear cell carcinoma patients [196]. We calculate the correlation between survival groups and clinical factors (Table 27). We found that survival groups are positively correlated to grade, stage, and metastasis, i.e. patients with higher grade, stage, metastasis, and lymphnode spread tend to have bad survival. However, correlation is not very high for any existing clinical factors. Hence, our genomic model provides a novel, reproducible, and objective means for predicting survival.

Table 27: Relationship of good and bad survival groups to other clinical factors.

Clinical Factor	**Pearson Correlation**
Grade (High vs. Low)	0.34 (p=0.000)
Stage (High vs. Low)	0.15 (p=0.012)
Metastasis	0.18 (p=0.002)
Lymphnode Spread	0.16 (p=0.065)

While training the prediction models, we perform 10 iterations of five-fold CV and select informative features based on a subset of training samples. In Table 28 and Table 29, we report histopathological properties and genes, respectively, which were selected in at least half of the CV models. Informative histopathological properties are similar to the ones observed in cancer grading such as nuclear shape and nuclear texture. Some of the genes in Table 29 have been linked to cancer. GEO profiles report differential (high) expression of C11orf75 in ER+ breast cancer as compared to ER- breast cancer

(GDS4056) [197]. MYLIP has been linked to progression in gastric cancer [198] and disease-free survival in renal clear cell carcinoma [199]. SLC17A4 is differentially expressed in low vs. high grade clear cell carcinoma [200]. Other genes reported in Table 29 can be useful biomarkers for kidney-cancer survival prediction, subject to further validation.

Table 28: Informative histopathological features associated with good and bad survival groups among KiCa patients.

Histopathological Feature	**Frequency of Selection**
Standard deviation in global graylevel intensity (bin 4)	0.88
Maximum nuclear minor axis length (bin 3)	0.84
Minimum no-stain shape elliptical area (bin 4)	0.78
GLCM information measure 2 of nuclear regions (bin 3)	0.78
Standard deviation in nuclear minor axis length (bin 3)	0.64
Standard deviation is green color distribution (bin 8)	0.56

Table 29: Informative genes associated with good and bad survival groups among KiCa patients.

Gene	Description	Frequency of Selection
C11orf75	Chromosome 11 open reading frame 75	0.9
WFDC3	WAP four-disulfide core domain 3	0.7
SLC10A7	Solute carrier family 10 (sodium/bile acid cotransporter family), member 7	0.68
LOC100133669	Uncharacterized	0.64
VWC2	Von Willebrand factor C domain containing 2	0.62
NLRP14	NLR family, pyrin domain containing 14	0.6
SUGT1P1	Suppressor of G2 allele of SKP1 (S. cerevisiae) pseudogene 1	0.58
TSKU	Tsukushi, small leucine rich proteoglycan	0.56
CARS	Cysteinyl-tRNA synthetase	0.52
NUDT19	Nudix (nucleoside diphosphate linked moiety X)-type motif 19	0.52
C3orf59	Mab-21 domain containing 2	0.5
MYLIP	Myosin regulatory light chain interacting protein	0.5
SLC17A4	Solute carrier family 17 (sodium phosphate), member 4	0.5

Ovarian Cancer Patient Stratification

Patient Stratification using Five-year Survival

In this section, we discuss the performance of the genomic survival-prediction model for OvCa patients trained using RNA-Seq data and clinical five-year survival information. Among train patients (Table 25), 82 and 24 patients have less than and more than five-year survival, respectively. We found that genomic data was not informative for survival prediction and model performed poorly with average 0.44 AUC for 10 iterations of five-fold CV on train set. When we applied the optimized model to stratify patients in the test set, we found that the survival functions for good and bad survival groups are not statistically significant. The p-value for one-sided logrank test on survival functions was 0.892 (Figure 77). One possible reason for this low performance could be that in this

chapter most OvCa samples (provided by TCGA) are grade-3 and stage III. Therefore, samples are histologically and genetically similar.

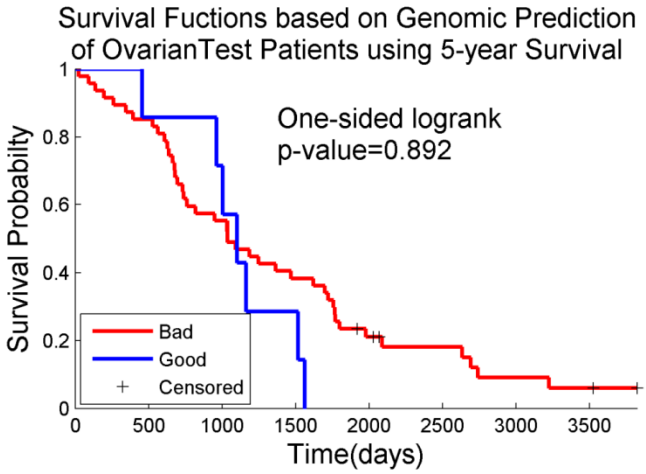

Figure 77: Survival functions of OvCa (test) patients based on a genomic prediction model trained using five-year survival information of train patients.
Survival functions are not separated and half of the time good survival function performs worse than bad survival function.

Correlation between Cluster Assessment Metrics and External-Validation Performance

Similar to kidney cancer, we train and test all genomic prediction models, calculate one-sided logrank p-value of survival functions, and calculate correlation of the p-value of test survival functions to all cluster-assessment metrics (Table 30). Unlike kidney cancer, histopathological clusters for ovarian cancer do not differ on genomic level and CV AUC was close to random performance in range 0.44 to 0.57. We found that none of metrics was able to measure biological usefulness of a clustering or correlate significantly in the expected direction.

Table 30: Correlation between cluster assessment metrics and testing performance.

Metrics based on Train Set	**Expected Direction of Correlation**	**Logrank p-value of survival functions based on Genomic Prediction of test patients**
Genomic CV AUC	Negative	0.522 (p=0.230)
Logrank p-value of Histopathology Clustering	Positive	-0.114 (p=0.808)
Adjusted Rand Index	Negative	-0.206 (p=0.657)
Rand Index	Negative	0.275 (p=0.550)
Hubert Index	Negative	0.275 (p=0.550)
Mirkin Index	Positive	-0.275 (p=0.550)
Global Silhouette Width	Negative	-0.184 (p=0.693)
Davies-Bouldin Index	Positive	0.100 (p=0.831)
Dunn's Index	Negative	-0.009 (p=0.985)

Limitations and Future Improvements

This study provides evidence that histopathological clustering can be informative for genomic patient stratification. However, current study has certain limitations including simple hierarchal clustering, genomic modeling for only two-cluster clusterings, and batch effects.

Hierarchal clustering often over-fits to the train data and the features are not reproducible in a separate dataset. Researchers have proposed consensus clustering methods, which cluster data multiple times on a subset of training samples and decide final clusters based on consensus to avoid over-fitting [201, 202].

Moreover, we have only trained and tested binary genomic models in this study. However, using PatientViz, we found several prognostically significant three- and four-cluster clusterings. In future work, we will like to study multi-class patient stratification

models. In addition, TCGA data were collected at multiple acquisition sources and suffer from batch effects. In this chapter, we have created train and test sets such that they are stratified by acquisition batches. However, in future we would remove batch effects using some normalization methods.

Conclusion

We developed and validated a novel method for generating prognostically significant patient stratification using genomic data. Proposed method first clustered patients based on their histopathological properties and then use the clustering labels for training genomic models. For KiCa, we select optimal clustering using two-sided logrank p-value of histopathological clustering and CV AUC of genomic model, which were found correlated with the model performance on the test set. We found genomic patient stratification model based on histopathological knowledge performs better than the model based on five-year survival information. Most selected genes for the genomic model include C11orf75, WFDC3, SLC10A7, LOC100133669, VWC2, NLRP14, SUGT1P1, TSKU, CARS, NUDT19, C3orf59, MYLIP, and SLC17A4. Some of these genes have been related to cancer while others could be potential biomarkers subject to validation. We have also designed a graphical user interface PatientViz that allows user to study cluster validity, reproducibility, and prognostic significance of different histopathological clustering generated using various parameter settings.

CHAPTER 11

CONCLUSIONS

This dissertation focused on developing imaging informatics algorithms for CDSSs based on histopathological WSIs. To achieve the three proposed specific aims, I developed the novel methods and the tools discussed in this dissertation. This chapter concludes the dissertation by summarizing the contributions to the field of pathology imaging informatics, followed by an outlook about the future directions of research and open challenges.

Contributions to Pathology Imaging Informatics

Connectivity-Based Threshold Estimation for Image Segmentation

In Chapter 2, I developed a robust segmentation method, ConnSoftT, which adaptively selects optimal segmentation thresholds for an image using a tissue connectivity model rather than color/intensity ranges. The selection of optimal segmentation thresholds is often challenging in applications with large variations in images. Quality control methods for tissue-fold artifact segmentation face similar challenge because of the differences in tissue morphology, tissue-fold thickness, and amount of stain among various WSIs. The proposed segmentation method adapts with the variations in tissue properties and in comparison to two other methods, ConnSoftT is more effective in detecting tissue-fold artifacts in 105 images each of OvCa and KiCa.

Batch-Invariant Color Segmentation

In Chapter 3, I developed a robust, supervised system for color segmentation of images. Supervised learning methods are often used for color segmentation applications, where the desired output is pre-defined color classes, such as histopathological stains. However, because of the differences in color properties of train (reference) and test (target) images, supervised methods often fail. I developed a supervised system that trains on multiple reference images (instead of a canonical reference) and normalizes colors in test images using a novel color map normalization. Color map normalization extracts color map of an image using unique colors and normalizes the distributions of individual color channels using non-parametric quantile normalization. Using semi-supervised color segmentation as ground truth, the proposed segmentation system performs with average pixel-level classification accuracy of 85% on 200 images from four batches.

Edge-based Nuclear Cluster Segmentation

In Chapter 4, I developed an edge-based method for segmenting complex structures (clusters) created by overlapping nuclei. In a diseased tissue, nuclei vary in size, shape, and texture properties. Therefore, modeling a single nucleus and then segmenting individual nuclei using region-based matching is very complex. Despite the shape and size differences, when two nuclei overlap a concavity is formed at the point of intersection of outer edges. The proposed edge-based method segments nuclear cluster into regions using neighboring concavities on a cluster edge. Thereafter, it fits ellipses on segmented regions and extracts nuclei. I found that the nuclear-shape features, extracted using segmented nuclei, were informative for KiCa grading.

Biologically Interpretable Shape-based Description

In chapter 6, I developed novel shape-based features for information extraction from images. Existing shape-based features are limited because they usually assume either non-parametric or over-simplified (e.g. elliptical) shape models and approximate an average shape for an image (rather than capturing the variation). The proposed shape-based features (1) model shapes using Fourier shape descriptors, which can model complex shape contours, and (2) represent the variation in shapes across the image rather than calculating an average shape. Moreover, the design of the features facilitate easy biological interpretation by highlighting shapes in the images that lead to informative shape features for an endpoint. In a multi-class renal subtype classification, the proposed features outperform or complement traditional image features. Also, informative shapes mimics the diagnostic criteria of pathologists.

Quantization-based Knowledge Modeling Methods for Images

In Chapter 9, I developed quantization methods for modeling knowledge in large images. Large images, such as WSIs, often have different ROIs with different high-level human interpretation. While making decisions manually, humans focus on a relevant ROI and assess that region to make decisions. In contrast, the proposed quantization methods extract information from all different regions, exploit the information to group regions into data-dependent groups, quantify the percent of different regions to represent images. In case-studies on kidney and ovarian carcinoma clinical diagnosis, when compared to naïve information combination, quantized representations improved decision making by up to 7% AUC. In addition, quantized methods were less sensitive to prior ROI selection as compared to the naïve method.

Statistically Over-Represented Feature Subsets for various endpoints

Existing CDSSs based on histopathological images use a tailored image feature set specific to a cancer endpoint. This dissertation developed a comprehensive image feature set (Table 6 and Table 16) that can model different cancer endpoints using data mining approaches. Results in Chapters 5, 7, and 9 indicated that the comprehensive set has high prediction performance on a variety of binary and multi-class cancer endpoints. Moreover, while optimizing decision models using the comprehensive set, some feature subsets were statistically over-represented. For endpoints—such as cancer grade and subtype—that are diagnosed using histopathology, emergent features were biologically interpretable. For instance, nuclear shape features are significantly over-represented in renal Fuhrman grading, which is based on nuclear morphology. On the other hand, for endpoints—such as stage, metastasis, lymphnode spread, and survival—that are not diagnosed using histopathology, I discovered novel imaging markers. For instance, color and nuclear texture features are statistically overrepresented in renal patient 5-year survival.

Normalization Methods for Batch-Invariant Decision Making

One of the main limitations of existing CDSSs is the unstable performance in clinical setting. This limitation mainly arises from the difference in data properties of train and test images caused by batch effects. Batch effects and their detrimental effects on CDSS performance has been overlooked by pathology imaging informatics researchers. Most CDSSs are validated on a dataset collected using one experimental setup. In chapter 7, a study on four separately acquired datasets indicates that data batch can be a larger source of variance in image features compared to biological factors such as grade and subtype. Among data-level and information-level normalization methods, I

found that information-level methods, especially ComBatN, was most effective in removing batch effects and enhancing prediction performance. When compared to prediction models with no normalization, ComBatN improves performance in 87% cases.

Genomic Patient Stratification Using Histopathology

The discovery of a prognostically significant stratification among cancer patients is useful for therapeutic decisions. Such stratification is often done on genomic-level using unsupervised genomic clustering. However, because of a large number of genes and limited patient samples, such clusters are often not reproducible. I developed a patient stratification method that clusters patients using histopathological features and then develops a supervised model for classifying the clustered patient groups using genomic features. I discovered that the prognostic significance of histopathological clustering and the cross-validation performance of genomic model on train patients are good metrics for selecting optimal genomic stratification model. The optimized genomic model using histopathological knowledge was able to separate good and bad survival groups in test dataset with p-value less than 0.05.

Future Directions

Deployment of CDSSs in a Clinical Setting

In this dissertation, I have developed and validated several methods required in a robust CDSS based on WSIs. In the future, research could be conducted to develop a device-independent CDSS with fast parallel processors for a pathology laboratory. The CDSS would be trained, optimized, and tested on the biopsy slides collected on daily basis. Based on results in Chapter 7, completely eliminating detrimental effects of batch

effects on the prediction models will be a significant challenge for the clinical deployment.

Patient Stratification

Most of this dissertation focused on supervised learning methods for clinical diagnosis. I only presented a proof-of-concept case-study on the genomic stratification of KiCa patients using histopathological clustering. Unsupervised patient clustering is a large subfield of machine learning with a lot of potential for novel discoveries. In the future, the current work would be extended by including different clustering techniques such as consensus clustering and self-organizing maps. Moreover, researchers need to develop novel metrics for evaluating clustering quality.

Integration of Imaging, Genomic, and Proteomic Data

Imaging informatics methods developed in this dissertation has allowed comprehensive and robust representation of WSIs. In addition to imaging, patients can also be represented using genomic and proteomic profiles. Integration of these sources can lead to a holistic cancer diagnosis platform. Researchers at Bio-MIBLAB are investigating both information- and decision-level integration of these resources. Preliminary results indicate that decision-level integration performs better for cancer grade and survival diagnosis. In the future, researchers could develop novel data integration techniques at information-, knowledge-, and decision-levels.

Closing Remarks

Despite of biological and technical variations, whole-slide images of cancer biopsies contain a large source of knowledge. Researchers working in the field of pathology informatics have to develop robust quality control, information extraction,

knowledge modeling, and prediction methods to make clinical decision support systems a reality. Automatic, batch-invariant, and comprehensive nature of my imaging informatics algorithms validated by biological interpretation of cancer endpoints have provided a deeper understanding of ovarian and renal carcinomas.

APPENDIX A: SELECTED PUBLICATIONS

The work discussed in this thesis is a compilation of several years of research that resulted in the following peer-reviewed journal publications, book chapter, and conference proceedings.

Journal Publications

1. **S. Kothari**, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image classification using biologically interpretable shape-based features," *BMC Med Imag*, vol. 13, p. 9, 2013.

2. **S. Kothari**, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology Imaging Informatics for Quantitative Analysis of Whole-Slide Images," *J Am Med Inform Assoc*, doi:10.1136/amiajnl-2012-001540, 2013.

3. **S. Kothari**, J. H. Phan, and M. D. Wang, "Eliminating tissue-fold artifacts in histopathological whole-slide images to improve cancer-grade prediction," *J Pathol Inform*, vol. 4, p. 22, 2013.

4. **S. Kothari**, J. H. Phan, T. H. Stokes, A. O. Osunkoya, A. N. Young, and M. D. Wang, "Removing batch effects from histopathological images for enhanced cancer diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, p. 1, 2013.

Book Chapter

1. T. H. Stokes, **S. Kothari**, C. W. Cheng, M. D. Wang "Review of quality control and analysis algorithms for tissue microarrays as biomarker validation tools," *Microarray Image and Data Analysis: Theory and Practice*, CRC Press, 2013, in press.

Conference Proceedings

1. **S. Kothari**, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, A. N. Young, and M. D. Wang, "Automatic batch-invariant color segmentation of histological cancer images," in Proc *IEEE Int Symp on Biomedical Imaging: From Nano to Macro*, 2011, pp. 657-660.

2. **S. Kothari**, Q. Chaudry, and M. D. Wang, "Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques," in *Proc. IEEE Int Symp on Biomedical Imaging: From Nano to Macro*, 2009, pp. 795-798.

3. **S. Kothari**, Q. Chaudry, and M. D. Wang, "Extraction of informative cell features by segmentation of densely clustered tissue images," in Proc *IEEE Eng Med Biol Soc.*, 2009, pp. 6706-6709.

4. **S. Kothari**, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image feature mining reveals emergent diagnostic properties for renal cancer," in Proc IEEE Int Conf on Bioinformatics and Biomedicine, 2011, pp. 422-425.

5. **S. Kothari**, J. H. Phan, A. O. Osunkoya, and M. D. Wang, "Biological interpretation of morphological patterns in histopathological whole-slide images," in proc. ACM Conf on Bioinformatics, Computational Biology and Biomedicine, 2012, pp. 218-225.

6. **S. Kothari**, J. H. Phan, and M. D. Wang, "Scale normalization of histopathological images for batch invariant cancer diagnostic models," in Proc *IEEE Eng Med Biol Soc.*, 2012, pp. 4406 - 4409.

7. J. H. Phan, A. Poruthoor, **S. Kothari**, and M. D. Wang, "Exploration of genomic, proteomic, and histopathological; image data integration for clinical prediction," in proc. Proc *IEEE China Summit and Int Conf on Signal and Information Processing*, 2013, in press.

8. R. A. Hoffman, **S. Kothari**, J. H. Phan, and M. D. Wang, "A high-resolution tile-based approach for classifying biological regions in whole-slide histopathological images," in proc. *IFMBE Int. Conf. on Health Informatics*, 2013, in press.

In Preparation

1. **S. Kothari**, J. H. Phan, and M. D. Wang, "Towards optimal representation of large whole-slide histopathological images for decision making," in preparation.

2. J. H. Phan, **S. Kothari**, P. Wu, and M. D. Wang, "Multi-Modal Predictive Modeling of Cancer Endpoints Using Genomic and Imaging Data," in preparation.

REFERENCES

[1] M. Peleg and S. Tu, "Decision support, knowledge representation and management in medicine," *Yearb Med Inform,* pp. 72-80, 2006.

[2] T. Liu, H. Peng, and X. Zhou, "Imaging informatics for personalised medicine: Applications and challenges," *Int J Funct Inform Personal Med,* vol. 2, pp. 125-135, 2009.

[3] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature,* vol. 455, pp. 1061-1068, 2008.

[4] S. Kothari, J. H. Phan, T. H. Stokes, and M. D. Wang, "Pathology imaging informatics for quantitative analysis of whole-slide images," *J Am Med Inform Assoc,* in press.

[5] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev Biomed Eng,* vol. 2, pp. 147-171, 2009.

[6] H. Fox, "Is h&e morphology coming to an end?," *J Clin Pathol,* vol. 53, pp. 38-40, 2000.

[7] M. Y. Gabril and G. M. Yousef, "Informatics for practicing anatomical pathologists: Marking a new era in pathology practice," *Mod Pathol,* vol. 23, pp. 349-358, 2010.

[8] A. Wetzel, "Computational aspects of pathology image classification and retrieval," *J Supercomput,* vol. 11, pp. 279-293, 1997.

[9] T. J. Fuchs and J. M. Buhmann, "Computational pathology: Challenges and promises for tissue analysis," *Comput Med Imaging Graph,* vol. 35, pp. 515-530, 2011.

[10] W. Amin, U. Chandran, V. Parwani Anil, and J. Becich Michael, "Biomedical informatics for anatomic pathology," in *Essentials of anatomic pathology,* L. Cheng and D. G. Bostwick, Eds., ed: Springer New York, 2011, pp. 469-480.

[11] E. T. Sadimin and D. J. Foran, "Pathology imaging informatics for clinical practice and investigative and translational research," *N Am J Med Sci (Boston)*, vol. 5, pp. 103-109, 2012.

[12] J. Ho, A. V. Parwani, D. M. Jukic, Y. Yagi, L. Anthony, and J. R. Gilbertson, "Use of whole slide imaging in surgical pathology quality assurance: Design and pilot validation studies," *Hum Pathol*, vol. 37, pp. 322-331, 2006.

[13] R. Dunkle, "Role of image informatics in accelerating drug discovery and development," *Drug Discovery*, vol. 4, pp. 75-82, 2003.

[14] J. Melamed, M. W. Datta, M. J. Becich, J. M. Orenstein, R. Dhir, S. Silver, *et al.*, "The cooperative prostate cancer tissue resource: A specimen and data resource for cancer researchers.," *Clin Cancer Res*, vol. 10, pp. 4614-4621, 2004.

[15] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, *et al.*, "Towards a knowledge-based human protein atlas," *Nat biotechnol*, vol. 28, pp. 1248-1250, 2010.

[16] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, *et al.*, "Review of the current state of whole slide imaging in pathology," *J Pathol Inform*, vol. 2, p. 36, 2011.

[17] S. Palokangas, J. Selinummi, and O. Yli-Harja, "Segmentation of folds in tissue section images," in *Conf Proc IEEE Eng Med Biol Soc*, 2007, pp. 5642-5645.

[18] P. A. Bautista and Y. Yagi, "Detection of tissue folds in whole slide images," in *Conf Proc IEEE Eng Med Biol Soc*, 2009, pp. 3669-3672.

[19] D. Gao, D. Padfield, J. Rittscher, and R. McKay, "Automated training data generation for microscopy focus classification," in *Med Image Comput Comput Assist Interv*, 2010, pp. 446-453.

[20] H. S. Wu, J. Murray, S. Morgello, M. I. Fiel, T. Schiano, T. Kalir, *et al.*, "Restoration of distorted colour microscopic images from transverse chromatic aberration of imperfect lenses," *J Microsc*, vol. 241, pp. 125-131, 2011.

[21] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, G. Xiaojun, *et al.*, "A method for normalizing histology slides for quantitative analysis," in *Proc IEEE Int Symp Biomed Imaging*, 2009, pp. 1107-1110.

[22] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, *et al.*, "Colour normalisation in digital histopathology images," in *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, 2009, pp. 100-111.

[23] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An integrated framework for supervised color-texture segmentation and cell splitting," *IEEE Trans Med Imaging*, vol. 30, pp. 1661-1677, 2011.

[24] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan, "Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," *Pattern Recognition*, vol. 42, pp. 1080-1092, 2009.

[25] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, "A boosted bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies," *IEEE Trans Biomed Eng*, vol. 59, pp. 1205-1218, 2010.

[26] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, *et al.*, "Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods," *PLoS ONE*, vol. 6, p. e17238, 2011.

[27] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, *et al.*, "A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data," *Pharmacogenomics J*, vol. 10, pp. 278-291, 2010.

[28] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Comput Biol Chem*, vol. 34, pp. 215-225, 2010.

[29] W. Hsu, L. Lee Mong, and J. Zhang, "Image mining: Trends and developments," *J Intell Inf Syst*, vol. 19, pp. 7-23, 2002.

[30] A. Tabesh, M. Teverovskiy, P. Ho-Yuen, V. P. Kumar, D. Verbel, A. Kotsianti, *et al.*, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *IEEE Trans on Med Imaging*, vol. 26, pp. 1366-1378, 2007.

[31] T. Fuchs, P. Wild, H. Moch, and J. Buhmann, "Computational pathology analysis of tissue microarrays predicts survival of renal clear cell carcinoma patients," in *Med Image Comput Comput Assist Interv*, 2008, pp. 1-8.

[32] M. Rahman, P. Bhattacharya, and B. C. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," *IEEE Trans Inf Technol Biomed,* vol. 11, pp. 58-69, 2007.

[33] L. Yang, O. Tuzel, W. Chen, P. Meer, G. Salaru, L. A. Goodell, *et al.*, "Pathminer: A web-based tool for computer-assisted diagnostics in pathology," *IEEE Trans Inf Technol Biomed*, vol. 13, pp. 291-299, 2009.

[34] V. Kovalev, A. Dmitruk, I. Safonau, M. Frydman, and S. Shelkovich, "A method for identification and visualization of histological image structures relevant to the cancer patient conditions," in *Computer analysis of images and patterns*. vol. 6854, P. Real, D. Diaz-Pernil, H. Molina-Abril, A. Berciano, and W. Kropatsch, Eds., ed: Springer Berlin / Heidelberg, 2011, pp. 460-468.

[35] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, *et al.*, "A methodological approach to the classification of dermoscopy images," *Comput Med Imaging Graph,* vol. 31, pp. 362-373, 2007.

[36] Q. Chaudry, S. Raza, A. Young, and M. Wang, "Automated renal cell carcinoma subtype classification using morphological, textural and wavelets based features," *J of Signal Processing Syst,* vol. 55, pp. 15-23, 2009.

[37] P. W. Huang and C. H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans Med Imaging*, vol. 28, pp. 1037-1050, 2009.

[38] K. Jafari-Khouzani and H. Soltanian-Zadeh, "Multiwavelet grading of pathological images of prostate," *IEEE Trans Biomed Eng*, vol. 50, pp. 697-704, 2003.

[39] O. Sertel, J. Kong, U. Catalyurek, G. Lozanski, J. Saltz, and M. Gurcan, "Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading," *J of Signal Processing Syst*, vol. 55, pp. 169-183, 2009.

[40] K. Jun, H. Shimada, K. Boyer, J. Saltz, and M. Gurcan, "Image analysis for automated assessment of grade of neuroblastic differentiation," in *Proc IEEE Int Symp Biomed Imaging*, 2007, pp. 61-64.

[41] C. Meurie, G. Lebrun, O. Lezoray, and A. Elmoataz, "A comparison of supervised pixels-based color image segmentation methods. Application in cancerology," *WSEAS Transactions on Computers*, vol. 2, pp. 739-44, 2003.

[42] K. Mao, P. Zhao, and P. Tan, "Supervised learning-based cell image segmentation for p53 immunohistochemistry," *IEEE Trans Biomed Eng*, vol. 53, pp. 1153-1163, 2006.

[43] P. Ranefalla, L. Egevadb, B. Nordina, and E. Bengtssona, "A new method for segmentation of colour images applied to immunohistochemically stained cell nuclei," *Anal Cell Pathol*, vol. 15, pp. 145-156, 1997.

[44] C. Gunduz-Demir, M. Kandemir, A. Tosun, and C. Sokmensuer, "Automatic segmentation of colon glands using object-graphs," *Med Image Anal*, vol. 14, pp. 1-12, 2010.

[45] J. P. Monaco, J. E. Tomaszewski, M. D. Feldman, I. Hagemann, M. Moradi, P. Mousavi, *et al.*, "High-throughput detection of prostate cancer in histological sections using probabilistic pairwise markov models," *Med Image Anal*, vol. 14, pp. 617-629, 2010.

[46] P. Thevenaz and M. Unser, "Snakuscules," *IEEE Trans Image Process*, vol. 17, pp. 585-593, 2008.

[47] O. Schmitt and M. Hasse, "Radial symmetries based decomposition of cell clusters in binary and gray level images," *Pattern Recognition*, vol. 41, pp. 1905-1923, 2008.

[48] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans biomed eng*, vol. 57, pp. 841-852, 2010.

[49] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, p. 1, 2004.

[50] L. Boucheron, "Object-and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer," PhD thesis, University of California, Santa Barbara, 2008.

[51] L. A. D. Cooper, K. Jun, D. A. Gutman, W. Fusheng, S. R. Cholleti, T. C. Pan, *et al.*, "An integrative approach for in silico glioma research," *IEEE Trans Biomed Eng,* vol. 57, pp. 2617-2621, 2010.

[52] M. Muthu Rama Krishnan, M. Pal, R. R. Paul, C. Chakraborty, J. Chatterjee, and A. K. Ray, "Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of oral submucous fibrosis," *J Med Syst,* vol. 36, pp. 1746-1756, 2012.

[53] C. Gunduz, B. Yener, and H. Gultekin S, "The cell graphs of cancer," *Bioinformatics,* vol. 20, pp. i145-i151, 2004.

[54] C. C. Bilgin, P. Bullough, G. E. Plopper, and B. Yener, "Ecm-aware cell-graph mining for bone tissue modeling and classification," *Data Min Knowl Discov,* vol. 20, pp. 416-438, 2009.

[55] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, *et al.*, "Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology," *IEEE Trans Biomed Eng,* vol. 57, pp. 642-653, 2010.

[56] J. Sudbø, R. Marcelpoil, and A. Reith, "New algorithms based on the voronoi diagram applied in a pilot study on normal mucosa and carcinomas," *Anal Cell Pathol,* vol. 21, pp. 71-86, 2000.

[57] J. Sudbo, A. Bankfalvi, M. Bryne, R. Marcelpoil, M. Boysen, J. Piffko, *et al.*, "Prognostic value of graph theory-based tissue architecture analysis in carcinomas of the tongue," *Lab Invest,* vol. 80, pp. 1881-1889, 2000.

[58] A. Cruz-Roa, J. C. Caicedo, and F. A. González, "Visual pattern mining in histology image collections using bag of features," *Artif Intell Med,* vol. 52, p. 91, 2011.

[59] S. Raza, R. Parry, R. Moffitt, A. Young, and M. Wang, "An analysis of scale and rotation invariance in the bag-of-features method for histopathological image classification," in *Med image comput comput assist interv 2011*. vol. 6893, G.

Fichtinger, A. Martel, and T. Peters, Eds., ed: Springer Berlin / Heidelberg, 2011, pp. 66-74.

[60] M. Rahman, S. Antani, and G. Thoma, "A learning-based similarity fusion and filtering approach for biomedical image retrieval using svm classification and relevance feedback," *IEEE Trans Inf Technol Biomed,* vol. 15, pp. 640-646, 2011.

[61] J. C. Caicedo, F. A. González, and E. Romero, "Content-based histopathology image retrieval using a kernel-based semantic annotation framework," *J Biomed Inform,* vol. 44, pp. 519-528, 2011.

[62] F. Yu and H. S. Ip Horace, "Semantic content analysis and annotation of histological images," *Comput Biol Med,* vol. 38, pp. 635-649, 2008.

[63] H. L. Tang, R. Hanka, and H. H. S. Ip, "Histological image retrieval based on semantic content analysis," *IEEE Trans Inf Technol Biomed,* vol. 7, pp. 26-36, 2003.

[64] A. M. Marchevsky, R. Dulbandzhyan, K. Seely, S. Carey, and R. G. Duncan, "Storage and distribution of pathology digital images using integrated web-based viewing systems," *Arch Pathol Lab Med,* vol. 126, pp. 533-539, 2002.

[65] H. Chang, G. V. Fontenay, J. Han, G. Cong, F. L. Baehner, J. W. Gray, *et al.*, "Morphometic analysis of tcga glioblastoma multiforme," *BMC Bioinformatics,* vol. 12, p. 484, 2011.

[66] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans Pattern Anal Mach Intell,* vol. 29, pp. 40-51, 2007.

[67] J. Won-Ki, J. Schneider, S. G. Turney, B. E. Faulkner-Jones, D. Meyer, R. Westermann, *et al.*, "Interactive histology of large-scale biomedical image stacks," *IEEE Trans Vis Comput Graph,* vol. 16, pp. 1386-1395, 2010.

[68] R. Zwönitzer, T. Kalinski, H. Hofmann, A. Roessner, and J. Bernarding, "Digital pathology: Dicom-conform draft, testbed, and first results," *Comput Methods Programs Biomed,* vol. 87, pp. 181-188, 2007.

[69] M. M. Triola and W. J. Holloway, "Enhanced virtual microscopy for collaborative education," *Bmc Med Educ,* vol. 11, 2011.

[70] D. Mayerich, L. Abbott, and B. McCormick, "Knife-edge scanning microscopy for imaging and reconstruction of three-dimensional anatomical structures of the mouse brain," *J Microsc,* vol. 231, pp. 134-43, Jul 2008.

[71] D. J. Foran, L. Yang, W. Chen, J. Hu, L. A. Goodell, M. Reiss, *et al.*, "Imageminer: A software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology," *J Am Med Inform Assoc,* vol. 18, pp. 403-415, 2011.

[72] D. Romo, E. Romero, and F. González, "Learning regions of interest from low level maps in virtual microscopy," *Diagn Pathol,* vol. 6 Suppl 1, p. S22, 2011.

[73] M. D. DiFranco, G. O'Hurley, E. W. Kay, R. W. Watson, and P. Cunningham, "Ensemble based system for whole-slide prostate cancer probability mapping using color texture features," *Comput Med Imaging Graph,* vol. 35, pp. 629-645, 2011.

[74] C. H. Huang, A. Veillard, L. Roux, N. Loménie, and D. Racoceanu, "Time-efficient sparse analysis of histopathological whole slide images," *Comput Med Imaging Graph,* vol. 35, pp. 579-591, 2011.

[75] V. Roullier, O. Lézoray, V. T. Ta, and A. Elmoataz, "Multi-resolution graph-based analysis of histopathological whole slide images: Application to mitotic cell extraction and visualization," *Comput Med Imaging Graph,* vol. 35, pp. 603-615, 2011.

[76] J. Herold, C. Loyek, and W. Nattkemper Tim, "Multivariate image mining," *WIREs Data Mining Knowl Discov,* vol. 1, p. 2, 2011.

[77] C. L. Liu, W. Prapong, Y. Natkunam, A. Alizadeh, K. Montgomery, C. B. Gilks, *et al.*, "Software tools for high-throughput analysis and archiving of immunohistochemistry staining data obtained with tissue microarrays," *Am J Pathol,* vol. 161, pp. 1557-1565, 2002.

[78] L. A. D. Cooper, K. Jun, W. Fusheng, T. Kurc, C. S. Moreno, D. J. Brat, *et al.*, "Morphological signatures and genomic correlates in glioblastoma," in *Proc IEEE Int Symp Biomed Imaging,* 2011, pp. 1624-1627.

[79] E. K. Lobenhofer, G. A. Boorman, K. L. Phillips, A. N. Heinloth, D. E. Malarkey, P. E. Blackshear, *et al.*, "Application of visualization tools to the analysis of histopathological data enhances biological insight and interpretation," *Toxicol Pathol,* vol. 34, pp. 921-928, 2006.

[80] B. Lessmann, T. W. Nattkemper, V. H. Hans, and A. Degenhard, "A method for linking computed image features to histological semantics in neuropathology," *J Biomed Inform,* vol. 40, pp. 631-641, 2007.

[81] J. R. Iglesias-Rozas and N. Hopf, "Histological heterogeneity of human glioblastomas investigated with an unsupervised neural network (som)," *Histol Histopathol,* vol. 20, pp. 351-356, 2005.

[82] I. Stephanakis, G. Anastassopoulos, and L. Iliadis, "Color segmentation using self-organizing feature maps (sofms) defined upon color and spatial image space," in *Artificial neural networks – icann 2010*. vol. 6352, K. Diamantaras, W. Duch, and L. Iliadis, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 500-510.

[83] M. Datar, D. Padfield, and H. Cline, "Color and texture based segmentation of molecular pathology images using hsoms," in *Proc IEEE Int Symp Biomed Imaging*, 2008, pp. 292-295.

[84] A. Rabinovich, S. Krajewski, M. Krajewska, A. Shabaik, S. M. Hewitt, S. Belongie, *et al.*, "Framework for parsing, visualizing and scoring tissue microarray images," *IEEE Trans Inf Technol Biomed,* vol. 10, pp. 209-219, 2006.

[85] L. Yang, W. Chen, P. Meer, G. Salaru, L. A. Goodell, V. Berstis, *et al.*, "Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens," *IEEE Trans Inf Technol Biomed,* vol. 13, pp. 636-644, 2009.

[86] R. Gutiérrez, F. Gómez, L. Roa-Peña, and E. Romero, "A supervised visual model for finding regions of interest in basal cell carcinoma images," *Diagn Pathol,* vol. 6, p. 26, 2011.

[87] K. Thomas, M. Sottile, and C. Salafia, "Unsupervised segmentation for inflammation detection in histopathology images," in *Image and signal processing*. vol. 6134, A. Elmoataz, O. Lezoray, F. Nouboud, D. Mammass, and J. Meunier, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 541-549.

[88] S. Kothari, J. H. Phan, A. O. Osunkoya, and M. D. Wang, "Biological interpretation of morphological patterns in histopathological whole-slide images," in *Proc ACM Conf on Bioinformatics, Computational Biology and Biomedicine*, 2012, pp. 218-225.

[89] S. Samsi, A. K. Krishnamurthy, and M. N. Gurcan, "An efficient computational framework for the analysis of whole slide images: Application to follicular lymphoma immunohistochemistry," *J Comput Sci*, vol. 3, pp. 269-279, 2012.

[90] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[91] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507-2517, 2007.

[92] M. A. Hall, "Correlation-based feature selection for machine learning," Department of Computer Science, Waikato University, New Zealand, 1999.

[93] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol*, vol. 3, pp. 185-205, 2005.

[94] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Machine learning: Ecml-94*. vol. 784, F. Bergadano and L. De Raedt, Eds., ed: Springer Berlin / Heidelberg, 1994, pp. 171-182.

[95] J. Kittler, "Feature set search algorithms," *Pattern recognition and signal processing*, pp. 41-60, 1978.

[96] D. B. Skalak, "Prototype and feature selection by sampling and random mutation hill climbing algorithms," in *Conf proc on machine learning*, 1994, pp. 293-301.

[97] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Ann Arbor: Michigan University, 1975.

[98] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, p. 671, 1983.

[99] S. Srivastava, J. J. Rodríguez, A. R. Rouse, M. A. Brewer, and A. F. Gmitro, "Computer-aided identification of ovarian cancer in confocal microendoscope images," *J Biomed Opt,* vol. 13, p. 024021, 2008.

[100] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[101] L. Breiman, "Random forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[102] L. van der Maaten, E. Postma, and J. van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research,* vol. 10, pp. 1-41, 2009.

[103] Z. Lei, A. W. Wetzel, J. Gilbertson, and M. J. Becich, "Design and analysis of a content-based pathology image retrieval system," *IEEE Trans Inf Technol Biomed,* vol. 7, pp. 249-255, 2003.

[104] F. Schnorrenberg, C. S. Pattichis, C. N. Schizas, and K. Kyriacou, "Content-based retrieval of breast cancer biopsy slides," *Technol Health Care,* vol. 8, pp. 291-297, 2000.

[105] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *Int J Med Inform,* vol. 77, pp. 81-97, 2008.

[106] S. Doyle, J. Monaco, M. Feldman, J. Tomaszewski, and A. Madabhushi, "An active learning based classification strategy for the minority class problem: Application to histopathology annotation," *BMC Bioinformatics,* vol. 12, p. 424, 2011.

[107] E. Cosatto, M. Miller, H. P. Graf, and J. S. Meyer, "Grading nuclear pleomorphism on histological micrographs," in *Proc Int Conf Pattern Recogn*, 2008, pp. 1-4.

[108] G. Begelman, M. Pechuk, E. Rivlin, and E. Sabo M D, "A microscopic telepathology system for multiresolution computer-aided diagnostics," *J Multimed,* vol. 1, 2006.

[109] S. Kothari, J. H. Phan, and M. D. Wang, "Eliminating tissue-fold artifacts in histopathological whole-slide images to improve cancer-grade prediction," *J Pathol Inform,* in press.

[110] L. A. Cooper, J. Kong, D. A. Gutman, F. Wang, J. Gao, C. Appin, *et al.*, "Integrated morphologic analysis for the identification and characterization of disease subtypes," *J Am Med Inform Assoc,* vol. 19, pp. 317-323, 2012.

[111] P. A. Bautista and Y. Yagi, "Improving the visualization and detection of tissue folds in whole slide images through color enhancement," *J Pathol Inform,* vol. 1, p. 25, 2010.

[112] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 929-944, 2007.

[113] X. Jiang, C. Marti, C. Irniger, and H. Bunke, "Distance measures for image segmentation evaluation," *EURASIP Journal on Advances in Signal Processing,* vol. 2006, p. 035909, 2006.

[114] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, *et al.*, "Automatic batch-invariant color segmentation of histological cancer images," in *Proc IEEE Int Symp Biomed Imaging*, 2011, pp. 657-660.

[115] S. Kothari, Q. Chaudry, and M. D. Wang, "Extraction of informative cell features by segmentation of densely clustered tissue images," in *Conf Proc IEEE Eng Med Biol Soc*, 2009, pp. 6706-6709.

[116] S. Kothari, Q. Chaudry, and M. D. Wang, "Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques," in *Proc IEEE Int Symp on Biomedical Imaging: From Nano to Macro*, 2009, pp. 795-798.

[117] C. Di Rubeto, A. Dempster, S. Khan, and B. Jarra, "Segmentation of blood images using morphological operators," in *Pattern Recognition, 2000 Proceedings 15th International Conference on*, 2000, pp. 397-400 vol.3.

[118] H. Refai, L. Li, T. K. Teague, and R. Naukam, "Automatic count of hepatocytes in microscopic images," in *Image Processing, 2003 ICIP 2003 Proceedings 2003 International Conference on*, 2003, pp. II-1101-4 vol.3.

[119] B. Nilsson and A. Heyden, "Segmentation of dense leukocyte clusters," in *Mathematical Methods in Biomedical Image Analysis, 2001 MMBIA 2001 IEEE Workshop on*, 2001, pp. 221-227.

[120] E. Glory, V. Meas-Yedid, G. Stamon, C. Pinset, and J. C. Olivo-Marin, "Automated image-based screening of cell cultures for cell therapy," in *Biomedical Imaging: Nano to Macro, 2006 3rd IEEE International Symposium on*, 2006, pp. 259-262.

[121] W. Weixing and H. Song, "Cell cluster image segmentation on form analysis," in *Natural Computation, 2007 ICNC 2007 Third International Conference on*, 2007, pp. 833-836.

[122] X. Bai, C. Sun, and F. Zhou, "Splitting touching cells based on concave points and ellipse fitting," *Pattern Recognition*, vol. 42, pp. 2434-2446, 11// 2009.

[123] A. Fitzgibbon, M. Pilu, and R. B. Fisher, "Direct least square fitting of ellipses," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, pp. 476-480, 1999.

[124] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang, "Histological image feature mining reveals emergent diagnostic properties for renal cancer," in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, 2011, pp. 422-425.

[125] 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 IEEE Int Symp Biomed Imaging*, 2008, pp. 496-499.

[126] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "Self-organizing map in matlab: The som toolbox," in *Proceedings of the Matlab DSP Conference*, 1999, pp. 16-17.

[127] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man and cybernetics*, vol. 3, pp. 610-621, 1973.

[128] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern recognition*, vol. 24, pp. 1167-1186, 1991.

[129] S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on gabor filters," *IEEE Trans Image Process,* vol. 11, pp. 1160-1167, 2002.

[130] A. Laine and J. Fan, "Texture classification by wavelet packet signatures," *IEEE Trans Pattern Anal Mach Intell,* vol. 15, pp. 1186-1191, 1993.

[131] V. Strela, P. N. Heller, G. Strang, P. Topiwala, and C. Heil, "The application of multiwavelet filterbanks to image processing," *IEEE Trans Image Process,* vol. 8, pp. 548-563, 1999.

[132] F. Kuhl and C. Giardina, "Elliptic fourier features of a closed contour," *Computer graphics and image processing,* vol. 18, pp. 236–258, 1982.

[133] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences,* vol. 98, p. 5116, 2001.

[134] H. Peng, F. 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,* pp. 1226-1238, 2005.

[135] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, *et al.*, "Gominer: A resource for biological interpretation of genomic and proteomic data," *Genome Biol,* vol. 4, p. R28, 2003.

[136] J. Eble, G. Sauter, J. Epstein, and I. Sesterhenn, *Pathology and genetics of tumours of the urinary system and male genital organs*: IARC press Lyon, 2004.

[137] S. Kothari, J. Phan, A. Young, and M. Wang, "Histological image classification using biologically interpretable shape-based features," *BMC Med Imaging,* vol. 13, pp. 1-17, 2013.

[138] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: A systematic survey," *Rensselaer Polytechnic Institute, Tech Rep,* 2005.

[139] Q. Chaudry, S. H. Raza, Y. Sharma, A. N. Young, and M. D. Wang, "Improving renal cell carcinoma classification by automatic region of interest selection," in *Proc IEEE Int Conf Bioinformatics Biomed*, 2008, pp. 1-6.

[140] S. Waheed, R. A. Moffitt, Q. Chaudry, A. N. Young, and M. D. Wang, "Computer aided histopathological classification of cancer subtypes," in *Proc IEEE Int Conf Bioinformatics Biomed*, 2007, pp. 503-508.

[141] H. J. Choi and H. K. Choi, "Grading of renal cell carcinoma by 3d morphological analysis of cell nuclei," *Computers in Biology and Medicine*, vol. 37, pp. 1334-1341, 2007.

[142] C. François, C. Moreno, J. Teitelbaum, G. Bigras, I. Salmon, A. Danguy, *et al.*, "Improving accuracy in the grading of renal cell carcinoma by combining the quantitative description of chromatin pattern with the quantitative determination of cell kinetic parameters," *Cytometry Part B: Clinical Cytometry*, vol. 42, pp. 18-26, 2000.

[143] S. H. Raza, Y. Sharma, Q. Chaudry, A. N. Young, and M. D. Wang, "Automated classification of renal cell carcinoma subtypes using scale invariant feature transform," in *Conf Proc IEEE Eng Med Biol Soc*, 2009, pp. 6687-6690.

[144] D. Lee, S. Antani, and L. Long, "Similarity measurement using polygon curve representation and fourier descriptors for shape-based vertebral image retrieval," in *SPIE Medical Imaging* 2003, pp. 1283-1291.

[145] R. Rangayyan, N. El-Faramawy, J. Desautels, and O. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans Med Imaging*, vol. 16, pp. 799-810, 1997.

[146] W. Cukierski, K. Nandy, P. Gudla, K. Meaburn, T. Misteli, D. Foran, *et al.*, "Ranked retrieval of segmented nuclei for objective assessment of cancer gene repositioning," *BMC Bioinformatics*, vol. 13, p. 232, 2012.

[147] D. Comaniciu and P. Meer, "Cell image segmentation for diagnostic pathology," in *Advanced algorithmic approaches to medical image segmentation*, J. S. Suri, S. K. Setarehdan, and S. Singh, Eds., ed: Springer London, 2002, pp. 541-558.

[148] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods," *Neuroimage*, vol. 21, pp. 46-57, 2004.

[149] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, 2005, pp. 26-33 vol. 1.

[150] E. Persoon and K. Fu, "Shape discrimination using fourier descriptors," *IEEE Transactions on systems, man and cybernetics*, vol. 7, pp. 170-179, 1977.

[151] W. Wong, F. Shih, and J. Liu, "Shape-based image retrieval using support vector machines, fourier descriptors and self-organizing maps," *Information Sciences*, vol. 177, pp. 1878-1891, 2007.

[152] R. McGill, J. Tukey, and W. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, pp. 12-16, 1978.

[153] K. Tae-Yun, C. Hyun-Ju, C. Soon-Joo, and C. Heung-Kook, "Study on texture analysis of renal cell carcinoma nuclei based on the fuhrman grading system," in *Enterprise networking and Computing in Healthcare Industry, 2005 HEALTHCOM 2005 Proceedings of 7th International Workshop on*, 2005, pp. 384-387.

[154] J. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification," *Advances in neural information processing systems*, vol. 12, pp. 547-553, 2000.

[155] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," 1992, pp. 144-152.

[156] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, p. 27, 2011.

[157] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*: Springer Verlag, 2009.

[158] Y. Sakai, S. Watanabe, and S. Matsukuma, "Chromophobe renal cell carcinoma showing oncocytoma-like hyalinized and edematous stroma: A case report and review of the literature," *Urologic oncology*, vol. 22, pp. 461-464, 2004.

[159] S. Kothari, J. H. Phan, and M. D. Wang, "Scale normalization of histopathological images for batch invariant cancer diagnostic models," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2012, pp. 4406-4409, 2012.

[160] S. Kothari, J. H. Phan, T. H. Stokes, A. O. Osunkoya, A. N. Young, and M. D. Wang, "Removing batch effects from histopathological images for enhanced cancer diagnosis," *IEEE Journal of Biomedical and Health Informatics,* in press.

[161] J. H. Phan, C. F. Quo, C. Cheng, and M. D. Wang, "Multiscale integration of -omic, imaging, and clinical data in biomedical informatics," *IEEE Reviews in Biomedical Engineering,* vol. 5, pp. 74-87, 2012.

[162] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics,* vol. 8, pp. 118-127, 2007.

[163] S. Aksoy and M. Haralick Robert, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognition Letters,* vol. 22, pp. 563-582, 2001.

[164] M. Boedigheimer, R. Wolfinger, M. Bass, P. Bushel, J. Chou, M. Cooper, *et al.*, "Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories," *BMC Genomics,* vol. 9, p. 285, 2008.

[165] R. Rocha, J. Vassallo, F. Soares, K. Miller, and H. Gobbi, "Digital slides: Present status of a tool for consultation, teaching, and quality control in pathology," *Pathol Res Pract,* vol. 205, pp. 735-741, 2009.

[166] C. Loyek, N. Rajpoot, M. Khan, and T. Nattkemper, "Bioimax: A web 2.0 approach for easy exploratory and collaborative access to multivariate bioimage data," *BMC Bioinformatics,* vol. 12, p. 297, 2011.

[167] J. Kölling, D. Langenkämper, S. Abouna, M. Khan, and T. W. Nattkemper, "Whide—a web tool for visual data mining colocation patterns in multivariate bioimages," *Bioinformatics,* vol. 28, pp. 1143-1150, April 15, 2012 2012.

[168] Y. Liu, Y. Sun, R. Broaddus, J. Liu, A. K. Sood, I. Shmulevich, *et al.*, "Integrated analysis of gene expression and tumor nuclear image profiles associated with

chemotherapy response in serous ovarian carcinoma," *PLoS One,* vol. 7, p. e36383, 2012.

[169] R. A. Soslow, G. Han, K. J. Park, K. Garg, N. Olvera, D. R. Spriggs, *et al.*, "Morphologic patterns associated with brca1 and brca2 genotype in ovarian carcinoma," *Mod Pathol,* vol. 25, pp. 625-636, 2012.

[170] J. Kong, D. Cooper L A, F. Wang, A. Gutman D, J. Gao, C. Chisolm, *et al.*, "Integrative, multimodal analysis of glioblastoma using tcga molecular data, pathology images, and clinical outcomes," *IEEE Trans Biomed Eng,* vol. 58, pp. 3469-3474, 2011.

[171] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics,* vol. 17, pp. S22-S29, 2001.

[172] S. Doyle and A. Madabhushi, "Consensus of ambiguity: Theory and application of active learning for biomedical image analysis," in *Pattern recognition in bioinformatics*. vol. 6282, ed: Springer Berlin / Heidelberg, 2010, pp. 313-324.

[173] K. J. Busam, C. R. Antonescu, A. A. Marghoob, K. S. Nehal, D. L. Sachs, J. Shia, *et al.*, "Histologic classification of tumor-infiltrating lymphocytes in primary cutaneous malignant melanoma. A study of interobserver agreement," *Am J Clin Pathol,* vol. 115, pp. 856-860, 2001.

[174] K. Y. Jen, J. L. Olson, S. Brodsky, X. J. Zhou, T. Nadasdy, and Z. G. Laszik, "Reliability of whole slide images as a diagnostic modality for renal allograft biopsies," *Hum Pathol,* vol. 44, pp. 888-894, // 2013.

[175] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, *et al.*, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Sci Transl Med,* vol. 3, p. 108ra113, // 2011.

[176] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *arXiv preprint arXiv:11013354,* 2011.

[177] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," *Computer Vision–ECCV 2006,* pp. 490-503, 2006.

[178] M. Aly, M. Munich, and P. Perona, "Indexing in large scale image collections: Scaling properties and benchmark," *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 418-425, 2011.

[179] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, pp. 1-8, 2007.

[180] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, p. 91, 2006.

[181] M. Pichler, G. C. Hutterer, T. F. Chromecki, J. Jesche, K. Kampel-Kettner, P. Rehak, *et al.*, "Histologic tumor necrosis is an independent prognostic indicator for clear cell and papillary renal cell carcinoma," *American Journal of Clinical Pathology*, vol. 137, pp. 283-289, 2012.

[182] J. Eble, G. Sauter, J. Epstein, and I. Sesterhenn, "Clear cell renal cell carcinoma," in *Pathology and genetics of tumours of the urinary system and male genital organs*, ed: IARC press Lyon, 2004, pp. 23-24.

[183] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics, 2013," *CA: A Cancer Journal for Clinicians*, vol. 63, pp. 11-30, 2013.

[184] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer Cell*, vol. 17, p. 98, 2010.

[185] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, *et al.*, "Integrative subtype discovery in glioblastoma using icluster," *PLoS One*, vol. 7, p. e35236, 2012.

[186] M. J. Fackler, C. B. Umbricht, D. Williams, P. Argani, L. A. Cruz, V. F. Merino, *et al.*, "Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence," *Cancer Res*, vol. 71, pp. 6195-6207, 2011.

[187] L. A. Gravendeel, M. C. Kouwenhoven, O. Gevaert, J. J. de Rooi, A. P. Stubbs, J. E. Duijm, *et al.*, "Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology," *Cancer Res*, vol. 69, pp. 9065-9072, 2009.

[188] C. Boutros Paul and B. Okey Allan, "Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data," *Briefings in Bioinformatics,* vol. 6, pp. 331-343, 2005.

[189] N. Garge, G. Page, A. Sprague, B. Gorman, and D. Allison, "Reproducible clusters from microarray research: Whither?," *BMC Bioinformatics,* vol. 6, p. S10, 2005.

[190] M. McShane Lisa, D. Radmacher Michael, B. Freidlin, R. Yu, M.-C. Li, and R. Simon, "Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data," *Bioinformatics,* vol. 18, pp. 1462-1469, 2002.

[191] N. Belacel, C. Wang, and M. Cupelovic-Culf, "Clustering: Unsupervised learning in large biological data," *Statistical Bioinformatics: A Guide for Life and Biomedical Science Researchers,* p. 89, 2010.

[192] G. Valentini, "Clusterv: A tool for assessing the reliability of clusters discovered in DNA microarray data," *Bioinformatics,* vol. 22, pp. 369-370, 2006.

[193] M. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of rna-seq data," *Genome Biology,* vol. 11, p. R25, 2010.

[194] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel unsupervised feature filtering of biological data," *Bioinformatics,* vol. 22, pp. e507-e513, 2006.

[195] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification,* vol. 2, pp. 193-218, 1985.

[196] I. Frank, M. L. Blute, J. C. Cheville, C. M. Lohse, A. L. Weaver, and H. Zincke, "An outcome prediction model for patients with clear cell renal cell carcinoma treated with radical nephrectomy based on tumor stage, size, grade and necrosis: The ssign score," *The Journal of Urology,* vol. 168, pp. 2395-2400, 12// 2002.

[197] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, *et al.*, "Ncbi geo: Archive for functional genomics data sets—update," *Nucleic Acids Research,* vol. 41, pp. D991-D995, 2013.

[198] T. Ueda, S. Volinia, H. Okumura, M. Shimizu, C. Taccioli, S. Rossi, *et al.*, "Relation between microrna expression and progression and prognosis of gastric

cancer: A microrna expression analysis," *The Lancet Oncology,* vol. 11, pp. 136-146, 2// 2010.

[199] D. Wuttig, S. Zastrow, S. Füssel, M. I. Toma, M. Meinhardt, K. Kalman, *et al.*, "Cd31, ednrb and tspan7 are promising prognostic markers in clear-cell renal cell carcinoma revealed by genome-wide expression analyses of primary tumors and metastases," *International Journal of Cancer,* vol. 131, pp. E693-E704, 2012.

[200] C. Battaglia, E. Mangano, S. Bicciato, F. Frascati, S. Nuzzo, V. Tinaglia, *et al.*, "Molecular portrait of clear cell renal cell carcinoma: An integrative analysis of gene expression and genomic copy number profiling," *Emerging Research and Treatments in Renal Cell Carcinoma,* pp. 23-56, 2012.

[201] R. Shen, S. Wang, and Q. Mo, "Sparse integrative clustering of multiple omics data sets," *Memorial Sloan-Kettering Cancer Center Department of Epidemiology and Biostatistics Working Paper Series,* vol. 24, 2012.

[202] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial, "Efficient algorithms for accurate hierarchical clustering of huge datasets: Tackling the entire protein space," *Bioinformatics,* vol. 24, pp. i41-i49, 2008.

VITA

Sonal Kothari was born in Jodhpur, India, the daughter of Dr. Gautam Chand Kothari and Pushpa Kothari. She received her bachelor's degree in electronics and communication engineering at Jai Narain Vyas University in India. As part of her undergraduate curriculum, she participated in research at the Indian Institute of Science and the Tata Institute of Fundamental Research. She earned a master's degree at Georgia Institute of Technology, where she began to pursue a doctorate in electrical and computer engineering while working in the Bio-Medical Informatics and Bioimaging Lab under the guidance of Professor May D. Wang. Her research interests include computer-aided diagnosis, imaging informatics, pattern recognition, and machine learning. She has been a teaching assistant for undergraduate and graduate level classes on medical image processing. When she is not working, she enjoys traveling, hiking, sketching, and reading.