Digital Pathology: Diagnostic Errors, Viewing Behavior and Image Characteristics

Ezgi Mercan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Linda Shapiro, Chair

Joann Elmore

Su-In Lee

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Digital Pathology: Diagnostic Errors,
Viewing Behavior and Image Characteristics

Ezgi Mercan

Chair of the Supervisory Committee:
Professor Linda Shapiro
Computer Science and Engineering

Whole slide imaging technologies provide a unique opportunity to collect and analyze large amounts of data on pathologists' interactions with the digital slide. In this work, we are studying the underlying causes of diagnostic errors in histopathology. Instead of focusing on the detection of invasive cancer, we consider the full-spectrum of diagnoses that a pathologist encounters during clinical practice and aim to study *misidentification* and *misinterpretation* errors that may cause *overdiagnoses* or *underdiagnoses*. To this end, we use the *digiPATH* dataset that consists of 240 breast biopsies with diagnoses ranging from benign to invasive cancer, the actions of pathologists recorded during their interpretations of the slides and the diagnostic regions associated with the final diagnoses they assigned. Our work consists of three parts: region of interest localization, diagnostic classification and viewing behavior analysis.

The first part of our work introduces a novel methodology to extract the diagnostically relevant regions of interest from pathologists' viewing behavior, and a computer vision model to detect these regions automatically on unseen images. Region of interest (ROI) localization provides us with a set of regions on the whole slide that either leads to the correct diagnosis or distracts the pathologists.

The largest portion of this thesis is devoted to the diagnostic classification problem. Starting with a tissue labeling, we developed features that describe the tissue composition of the image

and the structural changes. We first introduce two models for the semantic segmentation of the regions of interest into tissue labels. Then, we define two different feature sets that are constructed from the tissue label images. The first feature set consists of superpixel-label frequency and co-occurrence histograms, which are common image features. The second set of features are a sequence of histograms that together comprise the *structure feature*, a new kind of image feature defined for the first time in this work. Instead of attempting a four-class classification (benign, atypia, DCIS and invasive), we classify images one diagnosis at a time starting with invasive versus benign and ending with atypia versus DCIS. We show that the superpixel-label frequency and co-occurrence histograms work best for the classification of the invasive cases while the structure feature is more suitable for the benign, atypia and DCIS cases. with atypia versus DCIS. We show that the superpixel-label frequency and co-occurrence histograms work best for the classification of the invasive cases while the structure feature is more suitable for the benign, atypia and DCIS cases. The final part is an analysis of the pathologists' behavior on the whole slide images. We first analyze the relationship between the identification of the correct ROI and the diagnosis. We show that the higher overlap with the consensus ROI is correlated with a higher diagnostic accuracy. Then, we introduce novel measurements of interpretative patterns and identify two strategies used by the pathologists: scanning and drilling. We demonstrate that the interpretation strategy does not change the diagnostic accuracy but drilling is the more efficient option. Although it does not affect the diagnostic outcome, the interpretation strategy is correlated with the pathologists' characteristics like gender, age, experience and nervousness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful for the mentorship of my advisor, Linda Shapiro, who gave me the opportunity of my life by accepting me to her research group, and her life. You are the reason I survived my PhD. Your impact on my life as a researcher and as a person is immeasurable.

This work would be impossible without the support of my advisor Su-In Lee. Your example inspires me to become a better researcher.

I would like to express sincere appreciation to my collaborators Joann Elmore, Donald Weaver, Tad Brunyè, Mara Rendi, Jamen Bartlett, Selim Aksoy and Raymond Tse for their support and guidance, my co-authors Dilip Nagarkar and Sachin Mehta for their contributions to this body of work, and all my lab-mates for the feedback and support they offered every week for the last six years. Most importantly, the wonderful researchers and staff at Harborview Medical Center and Fred Hutchinson Cancer Research Center -Tom Morgan, Paul Frederick, Andrea Radick, Ross Lambert, Natalia Oster, and Hannah Shucard- made it all possible.

I am thankful to my dear friends, Bilge and Safiye, for the warm friendship I enjoyed for the last six years and hope to enjoy for many years to come. Their willingness to consume great amounts of coffee kept me sane. My survival was possible just because these people are in my life: Müge, my partner in crime - my home away from home, was always just a phone call away; Semih cried with me, laughed with me, danced with me and never asked why; Ender never stopped asking when I am graduating; Kıvanç watched many awful and awesome movies with me; and Gökçe, my person, kept the light and laughter alive with unexpected postcards. Thank you.

Erkan, my darling, your mere existence gives me strength.

Finally, my family deserves the highest accolade for braving the distance and supporting me through all the ups and downs. Mom, dad, sis - *I did it.*

DEDICATION

to my parents

and my sister

and her cat

Chapter 1

INTRODUCTION

Cancer is a disease with more than 200 types causing 585,720 estimated deaths and 1,665,540 estimated new cases in 2014 in the U.S. [46]. Each year, millions of people depend on pathologists to interpret the biopsy slides for an accurate diagnosis of a benign or malignant lesion. Yet, diagnostic errors are made by humans on a daily basis [34]. An *overdiagnosis* is defined as the detection of a disease that will never cause symptoms or death during a patient's lifetime. In statistical analysis, overdiagnoses are considered to be false positive results. If undetected, they may lead to unnecessary treatments that usually cause more harm to the patient considering the side effects of the cancer treatment. Overdiagnoses create an economical burden and affect patients psychologically and physically. An *underdiagnosis* is defined as the failure to recognize or correctly diagnose a disease. In statistics, underdiagnoses are false negative results. Underdiagnoses are perceived more dangerous than overdiagnoses, because they may cause delays in critical treatment.

The importance of early detection in cancer is well understood and has been emphasized for decades. Today, regular screenings for certain populations are recommended and conducted especially in developed countries. However, there is a growing concern in the medical community that the fear of underdiagnosing a patient leads to overdiagnosis and contributes to the ever-increasing numbers of cancer cases. Cancer is a disease of aging populations, thus, understanding the underlying causes of diagnostic errors is even more critical now, considering the aging world population.

Overdiagnosis is a complex problem due to multiple factors, including the systematic problems in the healthcare system that threatens medical experts with malpractice suits for every missed diagnosis, the new imaging systems with ever increasing sensitivity and the frequent mass screenings for large populations in which the majority are at a low risk for developing cancer [33]. Although the exact percentage of cases overdiagnosed is unknown, several studies demonstrated its preva-

lence in the breast cancer diagnosis [78, 82]. Our work on understanding the diagnostic errors and developing automated diagnosis systems is another step towards standardizing and improving the quality of diagnoses in clinical practice.

Whole slide imaging (WSI) technologies have revolutionized diagnostic medicine. With WSI, it is possible to create a digital image of the entire glass histopathology slide and view it using a virtual slide viewer. Virtual slides created novel opportunities in pathology: they can be archived indefinitely and can be sent over electronically to any distance. Virtual slides have been used for archiving in biorepositories and medical education at a national scale. Until the advent of virtual slides, the process of diagnosis has been studied on conventional light microscopes using video recordings, think-out-loud methods, questionnaires and interviews with pathologists. Due to the nature of these methods, these studies have been limited by the participant and sample size. Yet, virtual slides provide scientists with an unparalleled opportunity to study the human factors in medical diagnoses.

Despite their advantages, until recently, the FDA regulated WSI systems as medical devices posing the greatest risk to patients and limited their use in diagnostic medicine [37][90]. In a recent press release, the FDA announced that it would now allow the marketing of the first WSI system for digital pathology [1]. The announcement quoted the efficiency of the digital slides and noted that the performance of the pathologists interpreting surgical pathology cases in digital format was comparable to those made using glass slides. Our study based on digital slides is timely and critical, since the U.S. healthcare system will be going through a big shift towards digital pathology.

Our work focuses on identifying the causes underlying diagnostic errors by studying the viewing behaviors of pathologists combined with computational analysis of digital whole slides. In the first part, we are proposing new image features with the descriptive and discriminative power to classify a large spectrum of architectural abnormalities associated with pre-invasive lesions of the breast. We design structural image features that are based on color, texture and spatial arrangement of biological structures. The second part of this work focuses on analysis of scanning behavior and the association between scanning patterns, diagnostic errors and image characteristics. We propose new techniques to analyze the interactions of pathologists with the virtual slides in order

to discover specific scanning patterns that can lead to accurate and efficient diagnoses.

1.1 Research Objectives

- To develop a system that detects diagnostically relevant regions of interest from whole slide images (Chapter 4).

- To develop image features to describe the structural changes that lead to different diagnostic categories (Chapters 5 and 6).

- To analyze the effect of viewing behavior on the diagnostic decision making process (Chapters 7 and 8).

1.2 Thesis Outline

The rest of this thesis is organized as follows:

Chapter 2 summarizes related literature for the histopathological image analysis, whole slide image analysis and behavior analysis. Chapter 3 describes the digiPATH study and the whole slide images, diagnoses and tracking data collected that constitute our dataset in this work.

Chapters 4, 5 and 6 introduces the pieces of an automated diagnosis system for whole slide biopsies. Chapter 4 describes a novel analysis method for the tracking data that identifies diagnostically relevant regions and a model for detecting regions of interest in unseen images using computer vision techniques. In Chapter 5, we define the eight tissue labels to detect and describe the different biological structures. We then present two different approaches to segment the images into these tissue labels: a superpixel-based approach that uses color and texture features, and a deep-learning-based approach that uses raw pixel values to learn the segmentation. Chapter 6 uses the segmentation results from the previous chapter to build two image features: the superpixel-label frequency and co-occurrence histograms and the structure feature. We demonstrate the performance of both features using different segmentation results from Chapter 5 in a diagnostic classification task. An automated diagnosis support system can take a whole slide im-

age, can detect the regions of interest (Chapter 4), segment the ROI into tissue labels (Chapter 5) and diagnose each ROI (Chapter 6) to make a final diagnosis.

Chapters 7 and 8 explore some of the factors affecting the diagnostic accuracy. In Chapter 7, we analyze the effect of identifying the correct region of interest on the diagnostic accuracy. In Chapter 8, we analyze the viewing behavior of the participants using a novel approach to quantify their scanning. We evaluate the effect of the pathologists' demographics on the scanning pattern, as well as the effect of the scanning pattern on the diagnostic accuracy and efficiency.

Chapter 2

RELATED LITERATURE

2.1 Histopathological Image Analysis

Computer Aided Diagnosis (CAD) systems have been used in radiology for over a decade. Image analysis complements the interpretation of radiologists and helps reduce the inter-observer variability and diagnostic errors. Even before the advent of microscopes that can produce whole slide images (WSI) or virtual slides of the biopsy samples, the analysis of histopathological images has been a research interest for the computer vision community. However, with the introduction of WSIs, recent research aims to achieve a standard that can lead to clinical use as well as new discoveries about the biology of the disease.

Histopathological image analysis research tackles many problems related to diagnosis of the disease, including nucleus detection [23, 49, 50, 113, 114], region of interest identification [8, 41, 47], prediction of clinical variables (diagnosis [20, 27–30], grade [10, 52–55, 100], survival time [12, 24, 38]) and recently identification of genetic factors controlling tumor morphology (gene expression [24, 56], molecular subtypes [19, 24]). One of the major research directions in this area is to develop image features for different problems and image types. Commonly used image features can be categorized as pixel-level features, object-level features and semantic features.

Pixel-level features (low-level features) are extracted from all pixels in the image, and they are usually used as building blocks for more complex features. Most commonly used pixel-level features describe the color and texture of the pixel, sometimes including the surrounding pixels. Different color spaces have been used to describe the color of pixels including red-green-blue (RGB) [38, 57, 106], hue-saturation-value (HSV) [10, 28], Luv [28] and Lab [27, 28, 40, 54, 55]. Texture features are calculated by using the intensities of the surrounding pixels in comparison to the pixel in question. The most common texture features are the Haralick co-occurence matrix

(GLCM) [10, 28, 29], wavelet-based submatrices [52], Gabor filter responses [29, 30], grey level run-length matrices [28] and local binary patterns (LBP) [38, 71].

Object-level features depend on the image type and the specific problem being solved. Most problems define the object of interest as the nucleus [24, 28, 29, 56, 68, 122], while there are some efforts to include features extracted from stroma [12, 69, 122], cytoplasm [12], lymphocytes [122], necrosis [24] and other anatomical structures (e.g. layers, glands and ducts) [20, 40, 67, 75, 77] depending on the tissue type. Object-level features include but are not limited to size, shape and texture of the objects. There has been an effort to summarize the objects in an image using *architectural features* that describe the spatial arrangement of objects in the image. Architectural or topological features include several graph representations such as Delaunay triangulations, Voronoi diagrams and minimum spanning trees [10, 20, 29, 40, 47, 53].

Semantic features are interpretable high-level features that are designed for specific image type and problem. They build on pixel-level and object-level features. The most common semantic features include first-order statistics [12, 47] calculated for the set of pixels or objects, histograms [56, 65] and bags-of-words [71]. Semantic features are problem specific and computationally expensive in comparison to low-level or object-level features.

2.2 Whole Slide Image Analysis

WSIs are very large images containing a large number of biologically related structures, some of which are relevant to diagnostic decision. Most histopathological image analysis research has been done on analyzing the regions of interest (ROI) that contain diagnostically relevant structures after they have been identified by human experts. The marking of ROIs by experts is a time consuming task and prone to human error. Some researchers have addressed the problem of automatically *identifying ROIs* in WSIs by training supervised [8, 27, 41, 58, 91] and unsupervised [13, 94, 107] models. However, most models require expert labels to evaluate their results and do not facilitate pathologists' viewing behavior in their analysis.

Since WSIs are very large images (for example 10K $\times$ 10K pixels), most computer vision techniques using the whole image cannot be applied directly to high resolution WSIs. *Multi-scale*

(multi-resolution) image analysis [30, 47, 55, 94, 99] is a technique that makes use of the pyramid representation in which an image is subjected to repeated subsampling to obtain smaller versions that construct the levels of the image pyramid. Starting from the top level, which represents the subsampled image in the lowest resolution, the analysis moves to the next level (higher resolution) for only part of the image. In the case of whole-slide images, multi-scale analysis is used to achieve efficiency and overcome the memory requirements of large data. Additionally, multi-scale analysis mimics the behavior of the pathologist who starts from the lowest resolution and switches to higher resolutions (zooms in) when necessary, similar to switching to a higher magnification objective lens on a microscope as needed. *Multi-field-of-view* approaches [10], on the other hand, use the high-resolution WSI for all levels but creates levels by placing several grids on the image to produce different sized blocks. It parallels the field of view of a pathologist on a conventional microscope.

Another common theme in whole slide image analysis is *parallel processing* [19, 27, 55, 58, 94, 97] of image tiles for efficiency. These systems usually crop WSIs into smaller tiles (e.g. 512x512-pixel) and run the image processing algorithm on each tile. A post processing step to merge the results from all tiles and to handle tile artifacts follows. These techniques take advantage of the latests developments in hardware and software to speed up the processing time for large WSIs.

2.3 Viewing Behavior Analysis

The studies comparing virtual microscopy to conventional microscopy are crucial towards wide-spread adoption of virtual microscopy. In a comparison study, Randell *et al.* [88] evaluated two groups of pathologists using virtual and conventional microscopes. The study concluded that virtual microscopes are as accurate as conventional microscopes for diagnosis but the time required for diagnosis is longer on virtual microscopy. Onega *et al.* [83] conducted a survey with the participation of 200 pathologists with results that support the experimental findings of Randell *et al.*. The survey found out that WSIs are perceived as accurate diagnostically and useful for obtaining second opinions but too slow for interpretation therefore not preferred for clinical use by half of the participants.

WSI technology enabled researchers to study the cognitive process behind the diagnosis. Virtual slides have been used in analyzing the relationship between scanning patterns and the level of expertise [16, 60, 63, 64, 70], understanding the diagnostic errors made by pathologists [26, 110] and evaluating the effect of the context (abnormal or normal surroundings) in grading of biopsy slides [14]. The techniques used in scanning pattern analysis include user annotations and comments [26, 110], eye (gaze) tracking [60, 63, 64, 70], video recordings [26], mouse tracking [87] and special software that records user interactions [71, 87, 110].

The literature motivates further research in viewing behavior analysis on virtual slides. To our best knowledge, there is no extensive effort to integrate image analysis with viewing pattern analysis. Histopathological image analysis is a mature research area that can complete the studies towards understanding pathologists' scanning patterns and diagnostic errors.

Chapter 3

DATASET: DIGITAL PATHOLOGY PROJECT

The *Digital Pathology (digiPATH)* study aims to compare the accuracy and efficiency of pathologists interpretations of digital slides vs. glass slides. To this end, glass and digital slides of 240 breast biopsies were collected. A set of diagnoses and a mapping scheme for broader diagnostic categories were developed.

3.1 Case Selection and Whole Slide Images

240 biopsies were selected using a random sampling strategy from over 19,000 eligible cases in the Breast Cancer Surveillance Consortium registry archives in New Hampshire and Vermont. The cases had a range of diagnostic difficulty as judged by an expert panel and the research participants. Diagnostic categories atypia and DCIS are oversampled in comparison to national estimates in order to increase statistical confidence for studying disease categories with lower prevalence. The atypia and DCIS cases tend to be more challenging diagnostically. Other criteria taken into consideration during the case selection include the patient age, breast density and biopsy type. Development of the final set of 240 cases from an initial set of 370 was described in detail in [84]. Table 3.1 summarizes the properties of the final 240 cases used in the study.

The H&E stained glass slides were scanned using an iScan CoreoAu® in 40X magnification. A technician and an experienced breast pathologist reviewed each digital image, rescanning as needed to obtain the highest quality. The average image size for 240 whole slide images was 90,000×70,000 pixels.

The 240 cases were randomly assigned to one of four test sets, with stratification to achieve balance for the patient and case characteristics.

Table 3.1: Patient and case characteristics.

Patient and case characteristics	*# Cases*	*(%)*
Patient characteristics		
Age		
40–49 years	118	(49.2)
50–59	122	(50.8)
Breast density		
Almost entirely fat	13	(5.4)
Scattered fibroglandular densities	105	(43.8)
Heterogeneously dense	97	(40.4)
Extremely dense	25	(10.4)
Case characteristics		
Biopsy type		
Core needle biopsy	138	(57.5)
Excisional biopsy	102	(42.5)
Glass consensus diagnosis		
Benign	72	(30.0)
Atypia	72	(30.0)
DCIS	73	(30.4)
Invasive	23	(9.6)
Total	240	(100.0)

3.2 Diagnostic Categories

The cases span a wide spectrum of diagnoses ranging from benign proliferations to invasive cancer. A set of 14 diagnoses and 4 diagnostic categories were developed for the analysis. A histology form was developed to collect the diagnoses at the end of each assessment from the pathologists (See Appendix A.2). The development of the diagnostic categories and the mapping schemes were described in great detail in [6]. The final diagnostic categories and mappings are given in Table 3.2. Figure 3.1 shows some example images from digiPATH with different diagnostic categories illustrating the variability in visual features in pre-invasive lesions.

Two mapping schemes for diagnosis were used in the study: primary and alternative mappings. The differences between the two mappings are in the diagnoses of Flat Epithelial Atypia (FEA), Atypical Lobular Hyperplasia (ALH) and Lobular Carcinoma in-situ (LCIS).

Atypical Lobular Hyperplasia (ALH) and Lobular Carcinoma in Situ (LCIS)

ALH and LCIS are lobular lesions and have distinct histologies. The underlying biology, including altered expression of *E-cadherin* protein, is also different for lobular carcinomas. The majority of breast carcinomas are ductal, and the digiPATH study aims to study ductal carcinomas specifically. Because cases were not selected for a diagnosis of lobular in situ neoplasia, ALH and LCIS were mapped to the benign in the primary mapping. According to the primary mapping, if ALH or LCIS were diagnosed, the case was mapped to the other diagnoses also on the slide or was classified as benign if no other diagnoses were noted. The consensus ROIs based on the primary mapping ignores the regions of ALH and LCIS, and were marked for any other diagnoses present on the slide. However, the histology form filled out by the participants lists ALH under 'Atypical lesion' and LCIS under 'Carcinoma in situ'. The alternative mapping places ALH into the *atypia* category and LCIS into the *DCIS* category, following the convention of the histology form. Both the histology form and the alternative mapping reflect a categorization based on the risk factor in terms of prognosis.

Mapping ALH and LCIS to the benign category was done purely for analytical purposes on the

clinical study. For the image analysis work, it is more relevant to map ALH to atypia and LCIS to carcinoma-in-situ because the image features would more likely align with these categories.

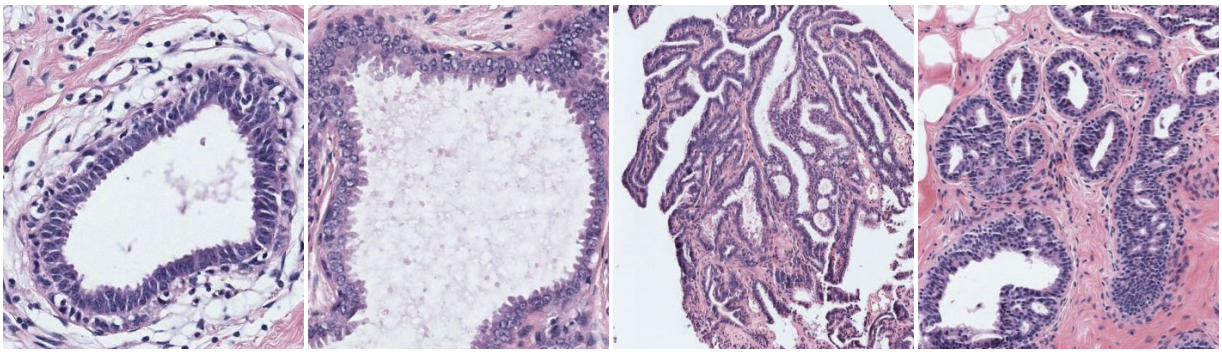
Flat Epithelial Atypa (FEA)

FEA is a neoplastic alteration of the ductal cells characterized by a low-grade cytological atypia. The atypical cells look flat, but cells are single layered or evenly stratified and there is no architectural atpia in the ductal structures. The majority of the atypia cases in the dataset involve a diagnosis of FEA as the most severe diagnostic finding in the expert consensus diagnosis. FEA was initially coded in the atypia category, but because FEA was not a diagnosis in our random selection strategy for the study sets, and because of its low risk for future carcinoma, it was later placed in the proliferative without atypia category in the primary mapping. The expert consensus ROIs based on the primary mapping ignores the regions of FEA because these lesions were not part of the primary clinical analysis, and were marked based on any other diagnoses present on the slide. However, it was placed under 'Atypical lesion' heading in the histology form, and the participant ROIs for the cases diagnosed as FEA would match the diagnosis FEA.

In this work, we are using the *alternative mapping*, because of the implied hierarchy in the histology form given to the participants. We assumed that when participants diagnosed a case as LCIS or ALH, they marked the regions exhibiting the characteristics of these diagnoses, not the other diagnoses that the case was mapped to according to its primary mapping. If we were to use the primary mapping, it would require us to remove all the cases with FEA, ALH and LCIS, as well as the participant assessments with these diagnosis from the analysis. Although we recognize the clinical complexity of FEA, ALH and LCIS; we chose to use the complete dataset in order increase the statistical power of our analyses.

Table 3.2: Mapping schemes for the diagnoses to the diagnostic categories.

Diagnosis	*Primary Mapping*	*Alternative Mapping*
Non-proliferative Changes Only	Benign	Benign
Fibroadenoma	Benign	Benign
Intraductal Papilloma Without Atypia	Benign	Benign
Usual Ductal Hyperplasia (UDH)	Benign	Benign
Columnar Cell Hyperplasia / Columnar Cell Change	Benign	Benign
Sclerosing Adenosis	Benign	Benign
Radial Scar/ Complex Sclerosing Lesion	Benign	Benign
Flat Epithelial Atypia (FEA)	Benign	Atypia
Atypical Ductal Hyperplasia (ADH)	Atypia	Atypia
Intraductal Papilloma With Atypia	Atypia	Atypia
Atypical Lobular Hyperplasia (ALH)	Benign	Atpia
Ductal Carcinoma in-situ (DCIS)	DCIS	DCIS
Lobular Carcinoma in-situ (LCIS)	Benign	DCIS
Invasive (ductal or lobular) cancer	Invasive	Invasive



(a) Columnar Cell Change

(b) Flat Epithelial Atypia

(c) Intraductal Papilloma without Atypia

(d) Usual Ductal Hyperplasia

(e) Intraductal Papilloma with Atypia

(f) Atypical Ductal Hyperplasia

(g) Ductal Carcinoma in-situ

(h) Invasive Carcinoma

Figure 3.1: Example diagnoses from digiPATH data.

3.3 Consensus Diagnosis and ROIs

240 glass slides were interpreted independently by three expert pathologists. Each expert pathologist filled out the histology form after each interpretation to list all the diagnoses observed on the slide. Subsequently, several in-person meetings and webinars were held to produce a glass consensus diagnosis for each case.

The process of individual interpretations and subsequent consensus meetings were repeated for the digital slides using a web-based virtual slide viewer, 9 months after the glass consensus meetings. In digital format, each expert also provided one or more regions of interest (ROI) for each case, supporting their final (mapping) diagnosis. The experts were allowed to draw ROIs in arbitrary sizes and numbers on each case. During the digital consensus meetings, one or more consensus ROIs were also determined for each digital slide. The process of the expert consensus review and the development of the diagnostic mapping were described in [6].

Digital consensus diagnosis differed from the glass consensus diagnosis for some cases; this analysis is presented in [34]. In this work, we will use the digital consensus diagnosis as our ground truth, since our work includes only digital slides. We collected an alternative version of the consensus ROIs from the expert panel to accommodate the alternative mapping and be consistent with the participant ROIs. Table 3.3 shows the distribution of the diagnostic classes in the dataset according to the digital consensus using the primary and alternative mappings.

Table 3.3: Distribution of the diagnostic categories based on digital consensus using the primary and alternative mappings.

Diagnostic Category	*Number of Cases (Primary Mapping)*	*Number of Cases (Alternative Mapping)*
Benign without atypia	72	60
Atypia	72	80
Ductal carcinoma in situ (DCIS)	73	78
Invasive breast cancer	23	22
Total	240	240

3.4 Participant Pathologist Data

200 pathologists from across the U.S. (Alaska, Maine, Minnesota, New Hampshire, New Mexico, Oregon, Vermont, and Washington) were invited to participate in the study. Participants were selected from pathologists who regularly interpret breast biopsy specimens in their clinical practices. Each participant was given a baseline survey about their demographics, experience and perceptions about breast pathology. Appendix A.1 includes a copy of the survey.

The study was designed as two phases (See Figure 3.2). In Phase I, half of the pathologists were randomized to interpret glass slides while the other half interpreted digital slides. In Phase II, after a 9-month washout period, half of the participants who interpreted glass slides in Phase I were randomized to interpret glass slides while the other half interpreted digital slides. Similarly, the participants who interpreted digital slides in Phase I were randomized to interpret glass or digital slides. From the participants who interpreted digital slides in Phase II, half were instructed to draw a rectangular region of interest at the end of their assessment. In this work, we are using the data collected in Phase II from the participants who interpreted digital slides.

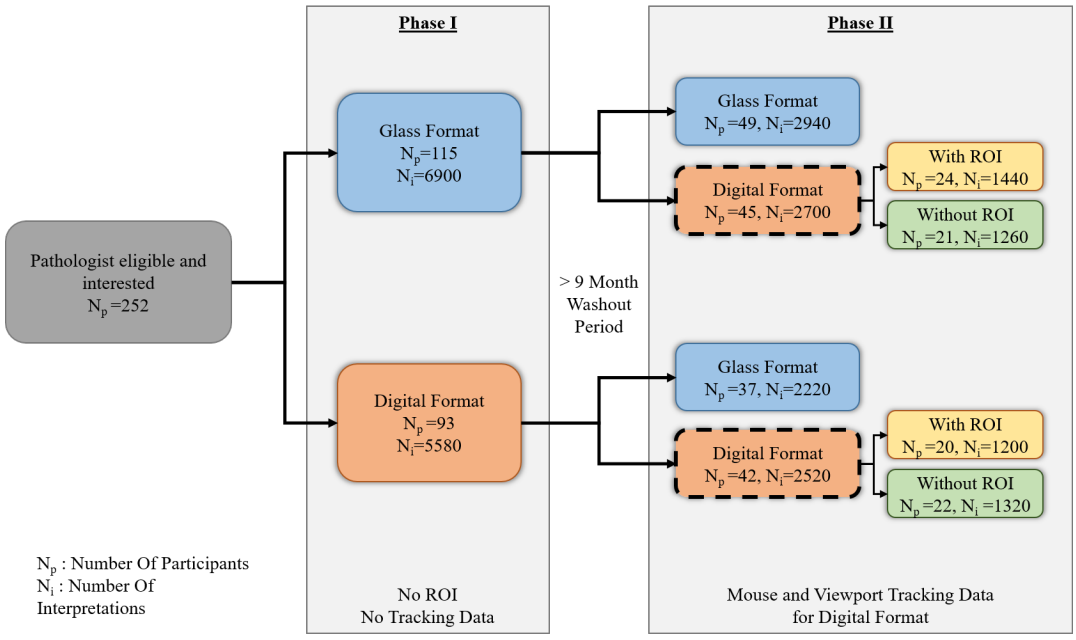


Figure 3.2: digiPATH study design

Participants interpreted a subset of the digital slides using a web-based virtual slide viewer. Each participant was randomly assigned a test set (N=60), and the order of the slides was also randomized. They filled out the histology form at the end of each interpretation. The histology form asked for their diagnoses, the perceived difficulty of the case and the participants' confidence in their diagnoses. Appendix A.2 includes a copy of the histology form. Half of the participants also provided an ROI supporting their diagnoses. This ROI was constrained to be a rectangle with a maximum size of 8603 pixels in height and width. Participants were instructed to include the area best supporting their diagnosis in the ROI. Figure 3.3 shows the process and the data collected from the participants who interpreted digital slides during Phase II.

Table 3.4 summarizes the data collected from the participants and experts.

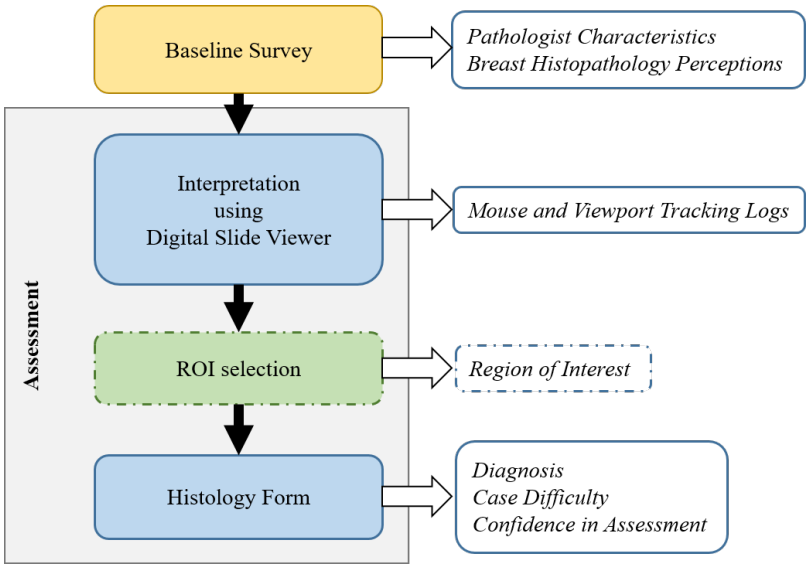


Figure 3.3: Participant data collected in Phase II in digital format

3.5 Viewport and Mouse Cursor Tracking Logs

One of the specific aims of digiPATH is to identify visual scanning patterns associated with diagnostic accuracy and efficiency. To this end, detailed viewport tracking and cursor tracking data were collected while pathologists interpreted digital slides using a web-based virtual slide viewer.

The software allows panning and zooming actions on high-resolution digital images and creates a log that records the changes occuring on the screen. Tracking data was collected for every interpretation made on the digital slides, including experts and all participants. Table 3.4 shows the summary of viewport and mouse cursor tracking data.

The *viewport tracking logs* are continuous streams of log entries where each entry describes a viewport (a rectangular part of the whole slide image visible on the participant's screen). A log entry was created whenever the image seen on the participant's screen changes. It includes the following variables: participant ID, case ID, log time (a date/time stamp in millisecond precision), coordinates of viewport (upper left corner, height and width), zoom level, participant's screen resolution (in pixels).

The *mouse cursor tracking logs* are also continuous streams of entries describing mouse cursor position on the participant's screen. An entry is created whenever the participant moves his mouse. It includes the following variables: participant ID, case ID, log time (a date/time stamp in millisecond precision), coordinates of the mouse cursor (x and y positions).

In addition to mouse cursor and viewport tracking data, a cognitive psychologist collected eye-tracking data to complement our visual scanning pattern research on digiPATH data [16]. Pilot studies showed that mouse cursor tracking is a good indicator of attention and the coupling between mouse cursor and eye-gaze positions is statistically significant [87]. Details of the eye-tracking study and analysis are given in [17, 18].

3.6 Summary

In this work, we used a subset of rich and extensive digiPATH data. As image set, digital WSIs of 240 breast biopsy specimens with the diagnostic categories benign, atypia, DCIS and invasive cancer were used. We opted for the digital consensus diagnosis and ROIs based on the alternative mapping scheme as the ground truth. Viewport tracking logs, ROIs and histology form data (including the diagnosis) from the participants were used to analyze the behavior, diagnostic accuracy and efficiency of the pathologists. Only half of the participants were asked to provide ROIs. Finally, participant demographics and breast pathology perceptions from baseline surveys were also

Table 3.4: Summary of the digiPATH data used in this work.

# Pathologists	*# Cases*	*# Interpretations*	*# ROI*	*Baseline Survey*	*Tracking Data*
1 digital consensus	240	-	428[a]	no	no
3 experts	240	720	730[a]	no	yes
44 participants	60	2640	2640	yes	yes
43 participants	60	2580	-	yes	yes

[a]Experts were allowed to draw more than one ROIs on each case.

included in the analysis. Table 3.4 summarizes the data.

Chapter 4

LOCALIZATION OF REGIONS OF INTEREST IN WHOLE SLIDE IMAGES

4.1 Introduction

4.1.1 Motivation

Whole slide images (WSI) have the potential to revolutionize histopathology by introducing computer aided support systems for medical experts. The diagnostic decision making is a complex cognitive process that includes visual search and interpretation tasks. A pathologist scans the whole slide image for diagnostically important areas, interprets each area and makes a decision based on the interpretations. For an automated diagnosis system, extraction of diagnostically relevant areas from whole slide images is a reasonable first step.

digiPATH data offers a unique opportunity in region of interest (ROI) localization, because we know exactly where each pathologist looked during his interpretation of the digital slides. The viewport logs give us a rich dataset that has not been studied before. In addition to the final diagnostic ROI, we know which parts of the image attract the pathologist's attention even though they are not marked by the pathologists. In addition to being the first step of automated diagnosis, identifying diagnostically important ROIs may help understand the pathologists' visual patterns of interpretation.

Our contribution in ROI localization is two-fold:

- We propose a novel methodology to extract ROIs from viewport logs of pathologists. Viewing logs are continuous streams of movements on a large WSI. By defining formal methods to extract ROIs from pathologist actions instead of asking pathologists to mark ROIs, we were able to use a large amount of data.

- We define the ROI localization problem as a classification task and show that using color and texture features, we can train a classifier to detect the ROIs in previously unseen images.

Methods and results presented in this chapter are previously published in [71] and [72].

4.1.2 Related Literature

The majority of the literature on whole slide images consider regions of interest marked by experts or the whole slide itself. In the case of TCGA images, the region of interest covers the whole slide, since the slides are prepared by taking sections from a tumor. Researchers working on datasets similar to TCGA data focused on segmenting the image into tumor and stroma, normalizing staining, removing imaging artifacts such as foldings, pen markings or background [58].

There are some studies that used the theories on human vision perception from cognitive science. Gutierrez *et al.* used functions inspired by human vision to combine over-segmented images and produce an activation map for relevant regions [41]. Their method is based on human perception of groupings, also known as Gestalt law. Using a supervised machine learning method, they merge relevant segments with the help of an energy function based their proximity to one another. They evaluate their findings with pathologist' drawn ROIs. Their method outperforms the standard saliency detection models.

Bahlmann *et al.* used a supervised model to detect ROIs using expert annotations [8]. They make use of color features to differentiate diagnostically relevant and irrelevant regions on a WSI. However, their evaluation considers only manually marked positive and negative samples and does not apply to the complete digital slide.

The experimental setting of Romo *et al* is the most relevant one to ours [91]. They use $70x70$ pixel tiles and calculate grayscale color histograms, LBP histograms, Tamura texture histograms and Sobel histograms for each tile. In a supervised setting, they classify all tiles in a WSI as diagnostically relevant or not. Their evaluation also makes use of actual interpretations of pathologists. They consider areas visited more as the ground truth positive and evaluate their results. Our methodology to extract ground truth ROIs is different and smarter than Romo *et al.*'s, since we

take all actions of the pathologists into account, not only duration.

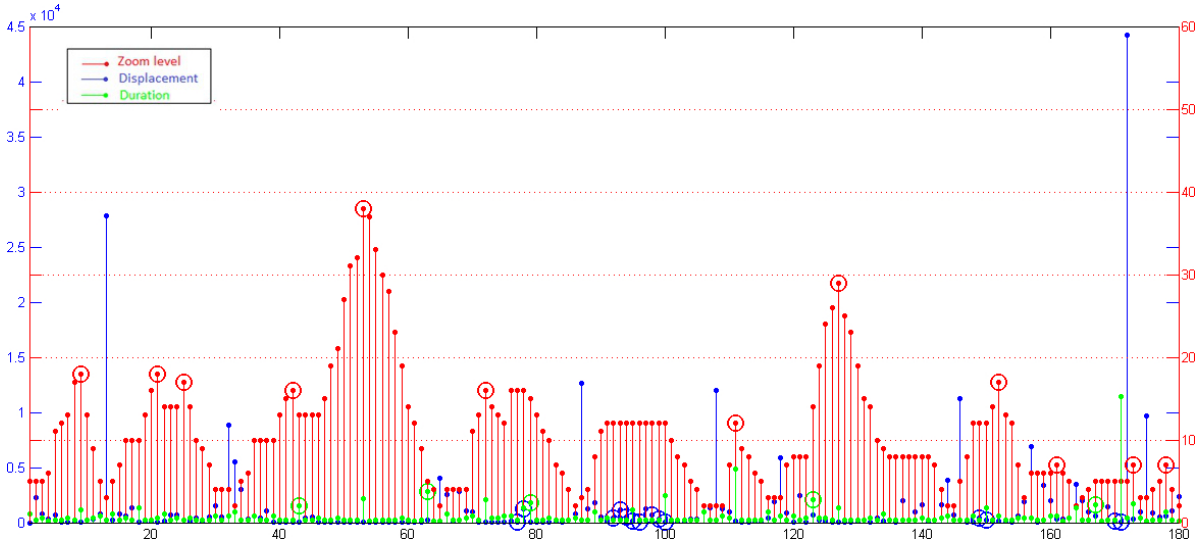
4.2 Region of Interest Extraction from Pathologists' Viewport Logs

Traditionally, regions of interest in whole slide images are provided by medical experts. Using one or two rectangular regions that support the final diagnosis of the WSI is an oversimplification of the complex diagnostic decision making process. In order to understand the underlying reasons for diagnostic errors and develop reliable diagnostic support systems, it is essential to mine more information than hand-marked ROIs from WSIs. In [71], we proposed a novel methodology that uses viewport tracking logs (see Section 3.5) to extract regions that a pathologist paid attention to. These regions include areas that support the final diagnosis as well as the areas pathologists focused during the interpretation, but decided not to include in the final assessment. Most importantly, these areas include distracting regions that lead the pathologists to make an incorrect diagnosis.

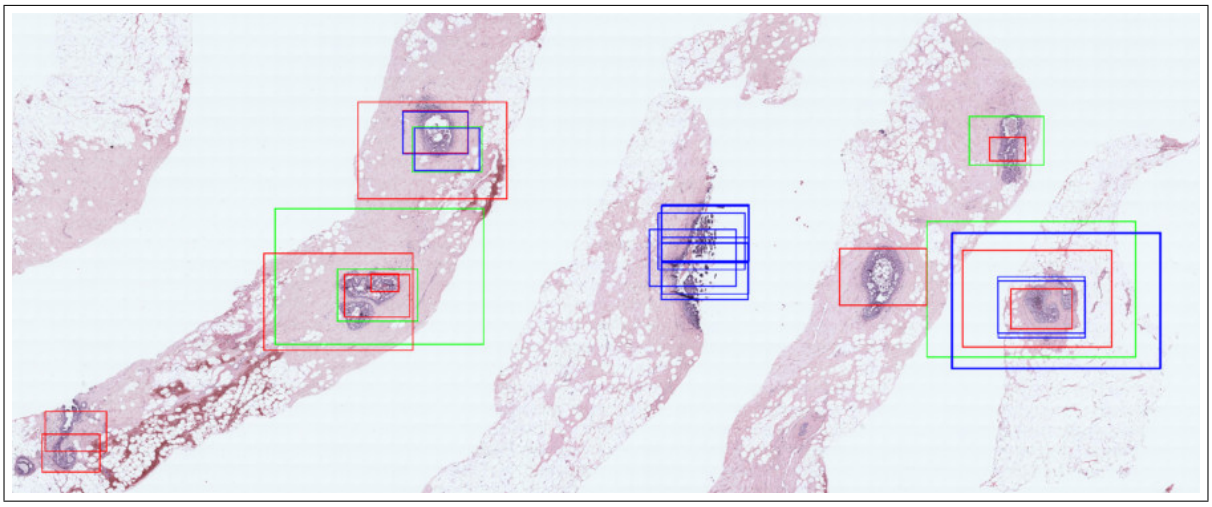
A viewport log is a stream of screen coordinates and zoom levels with timestamps that records the location of the pathologists screen in the digital whole slide. When combined, all viewports from a log can cover a large part of the whole slide, although we know there are special regions a pathologist focuses on in order to make a diagnosis. We used a graph to visualize a pathologists reading of a digital whole slide (see Figure 4.1 (a)) and defined three actions over the viewport tracking data that are used to extract regions where pathologist actually paid attention:

- Zoom peaks are the log entries where the zoom level is higher than the previous and the next entries. A zoom peak identifies a region where pathologist intentionally zoomed in to look in a higher resolution. See the circled red bars in Figure 4.1 (a). They are the local maxima points of the zoom level series plotted in red.

- Slow pannings are the log entries where the zoom level is the same with the previous entry and the displacement is small. We used a 100 pixel threshold on the screen level (100 x zoom on the actual image) to define slow pannings. The quick pans intended for moving the viewport to a far region result in high displacement values (more than 100 pixels). In



(a) Visualization of a Viewport Log



(b) Automatically extracted ROIs from the Viewport Log above

Figure 4.1: ROI Extraction from Viewport Logs: **(a)** An example visualization of the viewport log of a particular pathologist for the image in (b). The x-axis represents the log entry. The red bars represent the zoom level, the blue bars represent the displacement, and the green bars represent the duration. The zoom and duration values are shown on the vertical axis on the right of the figure, and the displacement values are shown on the vertical axis on the left of the figure. Note that the x-axis is not time; the actual time spent on each viewport is shown by the height of the green bars. The three types of selected actions are circled on the bars. **(b)** The rectangular regions visible on the pathologist's screen at the points selected from viewport log are drawn on the actual image. A *zoom peak* is a red circle in **(a)** and it corresponds to a red rectangle in **(b)**, a *slow panning* is a blue circle in **(a)** and it corresponds to a blue rectangle in **(b)**, a *fixation* is a green circle in **(a)** and it corresponds to a green rectangle in **(b)**.

comparison, slow pannings are intended for investigating a slightly larger area without completely moving the viewport. See the circled blue bars in Figure 4.1 (a). The zoom level represented by the red bars is constant and the replacement represented by blue bars is small at these points.

- Fixations are the log entries where the duration is longer than two seconds. Fixations identifies the areas to which a pathologist paid extra attention by looking at them longer. See circled green bars in Figure 4.1 (a).

After analyzing viewport tracking logs, the viewports that correspond to zoom peak, slow panning or fixation entries are marked as regions of interest. See Figure 4.1 (b) for example viewports: red rectangles are zoom peak viewports, blue rectangles are slow panning viewports and green rectangles are fixation viewports. Then, the union of these three actions are considered diagnostically important regions of interest.

4.3 Region of Interest Detection in Whole Slide Images

We use the visual bag-of-word (BoW) model [103] to represent the ROIs. The BoW model is a simple yet powerful representation technique that uses a predetermined dictionary to summarize documents (or images) with the frequencies of these words. In our framework, we used 120x120 pixel square patches as visual words and 3600x3600 sliding windows as bags. In selection of visual word and bag sizes, we considered the sizes of biological structures at 40X magnification. A visual word is constructed to contain more than one epithelial cell. A visual bag, on the other hand, may contain bigger structures such as breast ducts.

A visual vocabulary is a collection of distinct image patches that can be used to build images. The visual vocabulary is usually obtained by collecting all possible words (120x120 pixel patches) from all images and clustering them to reduce the number of distinct words. We selected two commonly used low-level image features for representing visual words: Local Binary Pattern (LBP) [115] histograms for texture and L*a*b* histograms for color. For LBP histograms, instead of using grayscale as is usually done, we used a well-known color deconvolution algorithm [46] to

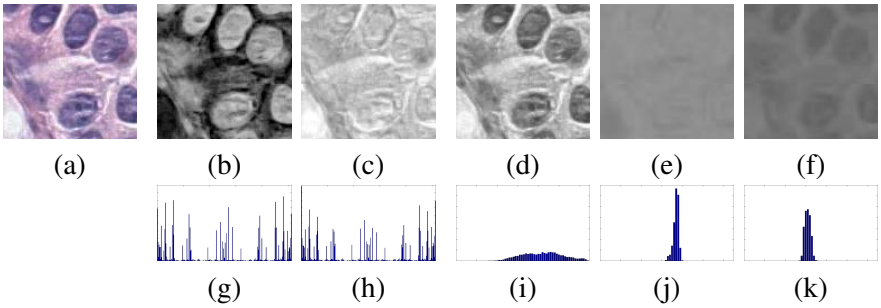


Figure 4.2: **Color and Texture Features: (a)** Original 120×120 pixel patch, **(b)** deconvolved color channel that shows good contrast for nuclei dyed with haematoxylin, **(c)** deconvolved color channel that shows good contrast for eosin dye, **(d, e, f)** L, a and b channels of the image in L*a*b* color space, **(g)**LBP histogram calculated on haematoxylin channel, **(h)** LBP histogram calculated on eosin channel, **(i, j, k)** color histograms of L, a and b channels. At the end, LBP histograms on two channels are concatenated to produce the first set of features, color histograms on L*a*b* channels are concatenated to produce the second set of features. ©2014, IEEE

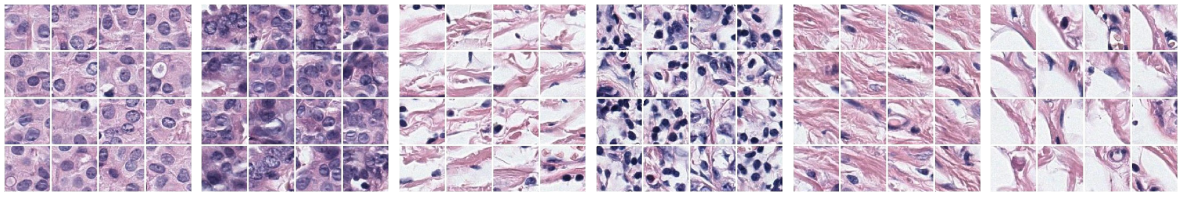
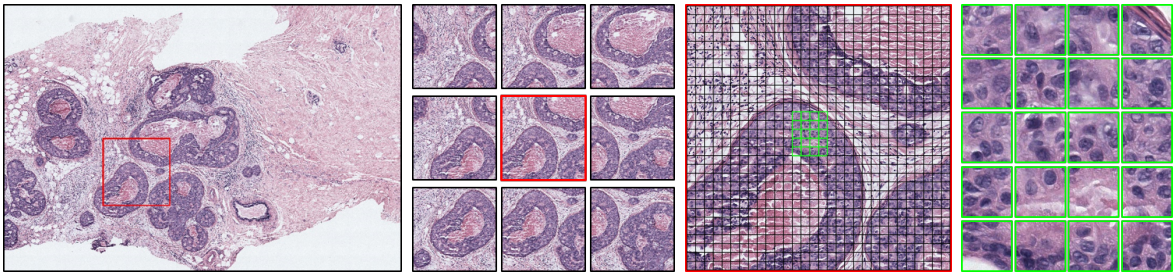


Figure 4.3: Example results from the K-means clustering. Each set shows 16 example patches from a cluster produced by the K-means algorithm with 100 clusters on LBP+L*a*b* features. Each cluster corresponds to a visual word and will be used to build visual vocabulary. Note that the example patches from the same cluster shows similar texture and color characteristics. ©2014, IEEE

obtain two chemical dye channels, haematoxylin and eosin (H&E) and calculated the LBP feature on these two color channels. For example images of RGB to H&E conversion, see Figure 4.2. Each visual word is represented by a feature vector that is the concatenation of the LBP and L*a*b* histograms. Both LBP and L*a*b* features have values ranging from 0 to 255, and we used 64 bins for each color and texture channel resulting in a feature vector of length 320.

We used k-means clustering to obtain a visual vocabulary that can be represented as the cluster centers. Any 120 x 120 pixel image patch is assigned to the most similar visual word, the one with the smallest Euclidean distance between the feature vector of the patch and the cluster center that



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

Figure 4.4: Sliding window and visual bag-of-words approach: **(a)** A 3600 3600 pixel sliding window is shown with a red square on an image region. **(b)** The sliding window from a is shown in the center with neighboring sliding windows overlapping 1200 pixels horizontally and vertically. **(c)** 120 $times$ 120 pixel visual words are shown with black borders on the same sliding window from (a). Visual words do not overlap. **(d)** A group of visual words are shown in higher magnification. They are identified with green borders in (c) ©2016, Journal of Digital Imaging

represents the visual word. This enables us to represent image windows (bags) as histograms of visual words. See Figure 4.3 for some example clusters. Since the cluster center is not always a sample point in the feature space, we show the closest 16 image patches to cluster centers for 6 visual words.

We used a sliding window approach for extracting visual bags that are 3,600 x 3,600 pixel image windows overlapping by 2,400 pixels both horizontally and vertically. Overlapping the sliding windows is a common technique to ensure that at least one window contains an object, if all others fail to encompass it. We picked a 2/3 overlap between sliding windows for performance purposes, since a higher overlap would increase the number of sliding windows and hence the sample size for the classification. Each sliding window contains 30 x 30=900 image patches, which are then represented as color and texture histograms and assigned to visual words by calculating distances to cluster centers. In this framework, each sliding window is represented as a histogram of visual words. Figure 4.4 shows an example sliding window and visual words computed from it.

4.4 Experiments and Results

We formulated the detection of diagnostically relevant ROIs as a binary classification problem where the samples are sliding windows [72]. Using visual bag-of-words histograms and the labels obtained through viewport analysis, we trained classifiers to label sliding windows as diagnostically relevant or not. Sliding windows overlapping with the ground truth ROIs are used as positive examples and we employed 10-fold cross-validation experiments using logistic regression.

We explored the effect of certain variables in the classification setting:

- Dictionary size in the visual bag-of-words model: We identified the visual words that are critical to describe diagnostically relevant ROIs.
- Visual word definitions: We assessed the performance of superpixels as visual words instead of square patches.
- Training data: We compared the hand-marked ROIs to the ROIs automatically detected from pathologist viewport logs.

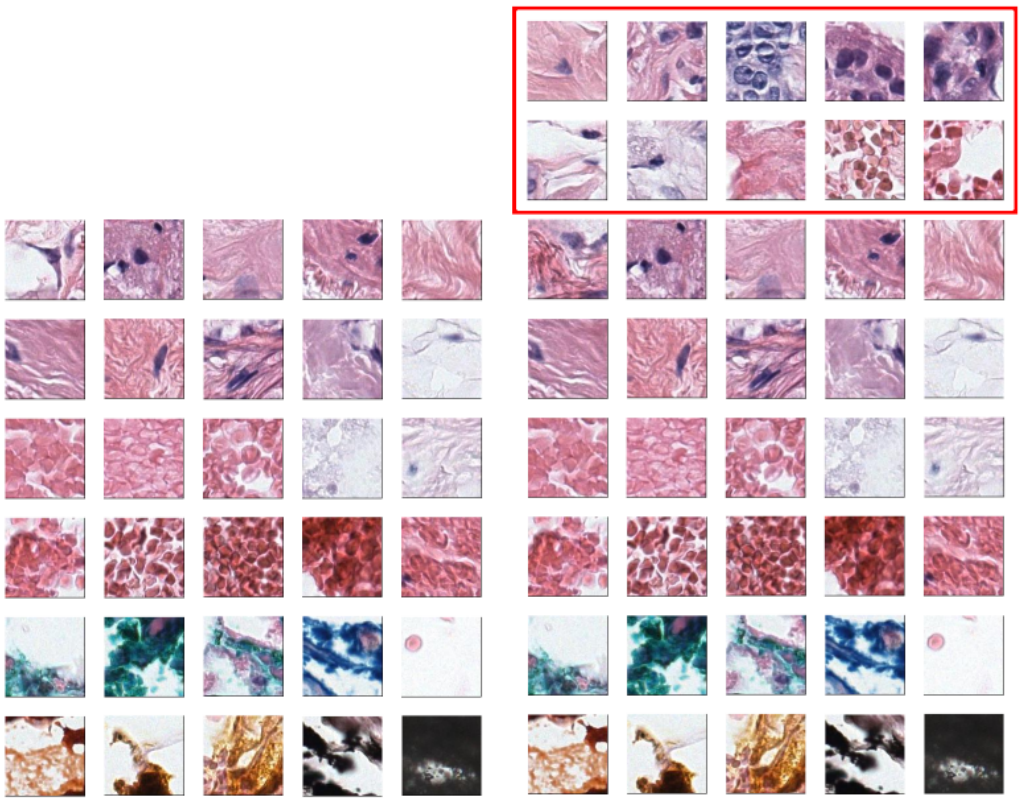
We used classification accuracy metric that is is calculated as the percentage of sliding windows that are correctly classified as diagnostically relevant or not over all possible sliding windows of size 3,600 x 3,600 pixels.

4.4.1 Dataset

We used 240 WSIs from the digiPATH dataset (Section 3.1), three expert pathologists' ROIs (Section 3.3) and three expert pathologists' viewport tracking logs (Section 3.5).

Effect of Dictionary Size on ROI Detection Accuracy

The dictionary size corresponds to the number of clusters and the length of the feature vector (as the histogram of visual words) calculated for each image window. For this reason, the dictionary size can determine the representative power of the model, yet large dictionaries present a computational



(a) (b)

Figure 4.5: Visual dictionaries with (b) 40 words and (a) 30 words. Note that visual words that represents epithelial cells are missing in (a) while present in (b). This difference causes classification accuracy to drop from 74% to 46%. The visual words that represent epithelial cells are absolutely necessary for the diagnostically relevant regions, since all the structures in (a) are discarded by pathologists during the screening process. ©2016, Journal of Digital Imaging

challenge and introduce redundancy. Since the dictionary is built in an unsupervised manner, we tested different visual vocabulary sizes to understand the effect of dictionary size on model predictions. For this purpose, we applied k-means clustering to obtain the initial 200 clusters from millions of image patches and reduced the number of clusters by using hierarchical clustering.

The classification accuracy (74%) does not change when the dictionary size is reduced from 200 to 40 but drops from 74% to 46% when the dictionary contains only 30 words. This trend is present in all experiments with different visual words (superpixels) and different training data. We compared the visual dictionaries with 40 words and 30 words to discover critical visual words

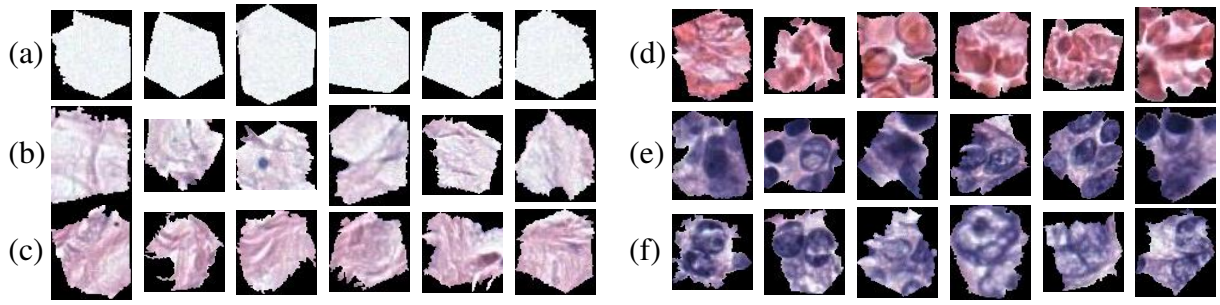


Figure 4.6: **Superpixel clusters:** Some superpixel clusters as visual words from a dictionary of superpixels. Most superpixel clusters can be named by expert pathologists although they are discovered through unsupervised methods. Some of the superpixel clusters as identified by pathologists: **(a)** empty space, **(b)** loose stroma, **(c)** stroma **(d)** blood cells, **(e)** epithelial nuclei, **(f)** abnormal epithelial nuclei.

in the ROI representation. Figure 4.5 shows the visual dictionaries; the missing words in the 30-word vocabulary are framed in the 40-word dictionary. The missing words include some blood cells, stroma with fibroblast and, in particular, epithelial cells in which the ductal carcinoma or pre-invasive lesions present abnormal features.

4.4.2 Superpixels vs Patches for ROI Detection

Superpixel [89] segmentation is a very popular method in computer vision. There has been successful work in histopathological image analysis in which superpixels are used as building blocks of the tissue analysis [12, 13]. To improve the visual word definition, we replaced 120x120 pixel image patches with superpixels that are obtained by the SLIC algorithm [3]. Similar to image patches, we calculated color and texture features from all superpixels from all images and built our visual vocabulary by k-means clustering. Figure 4.6 shows the closest 6 superpixels to cluster centers.

Superpixel segmentation is implemented an optimization problem that is computationally expensive to calculate. Using superpixels instead of square patches did not improve diagnostically relevant ROIs detection significantly. Figure 4.7 gives a comparison of ROI classification accuracy for superpixel-based visual words and square patch visual words.

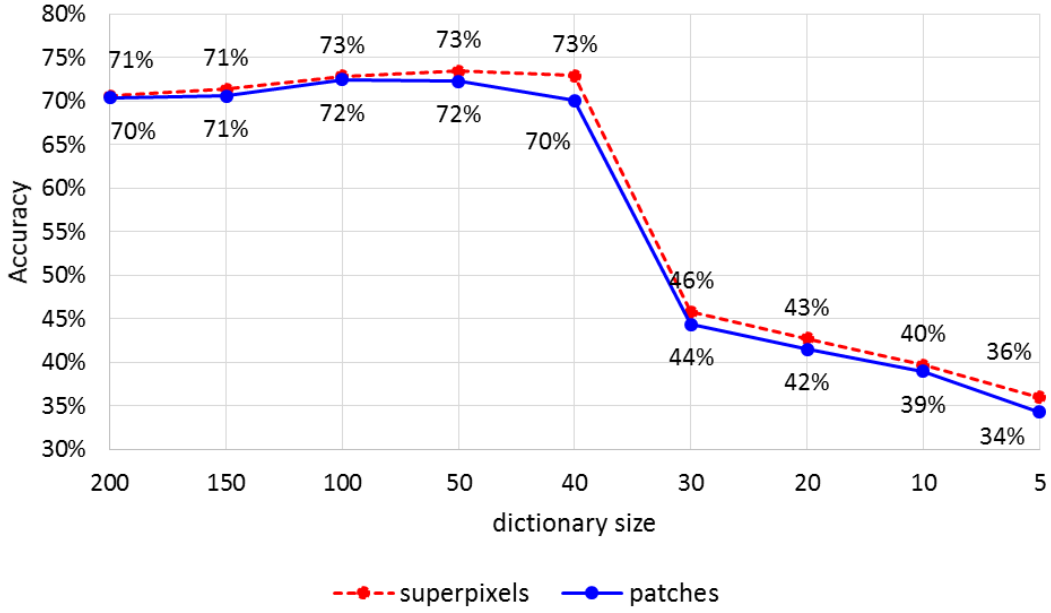
4.4.3 Manually Marked ROIs vs Automatically Extracted ROIs

The viewport analysis extracts a set of ROIs from the WSIs using only the zooming and panning actions of the pathologists. These are potentially diagnostically relevant even though not included in the diagnostic ROI that is drawn by the pathologist. However, due to the nature of viewing software or human factors, some of these areas are incorrectly depicted as diagnostically relevant because their zoom, duration or displacement characteristics are matched to our criteria. This situation introduces error in the training data with some false positives.

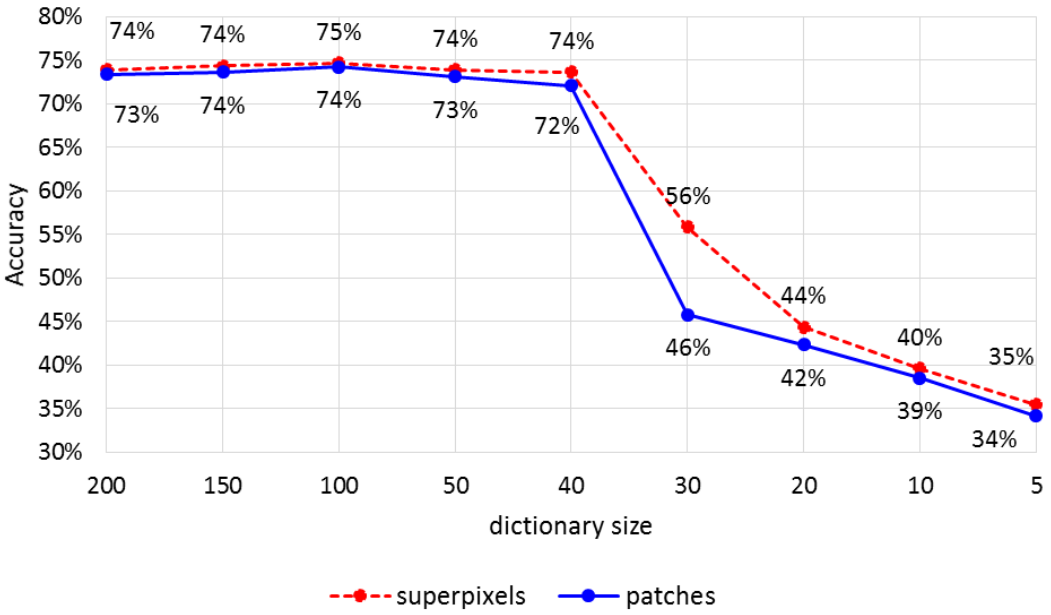
Alternative to the viewport analysis, we retrained our model by with the consensus ROIs that were agreed upon by three experts. Consensus ROIs show diagnostic features specific to the diagnosis of the slide but they are comparatively very small and very expensive to collect. Although the consensus ROIs are very controlled in nature and gold standard in research, they improve detection accuracy very little. Figure 4.7(a) shows that the classification accuracy for manually marked ROIs is only slightly higher to that of viewport ROIs as shown in Figure 4.7(b). In comparison, automated extraction of ROIs from viewport logs requires little to no effort from the pathologists and enables us to collect large amounts of ROIs.

4.4.4 Comparison of Predicted ROIs to ROIs Extracted from Viewports

We evaluated our ROI detection framework in a classification setting where each instance is an image region extracted by the sliding window approach. In addition to these quantitative evaluations, we produced probability maps that show the regions detected as ROIs by the computer. Figure 4.8 shows a comparison of viewport-extracted ROIs (ground truth) and predictions of the two different models. A visual evaluation reveals that our detection accuracy is affected by the rectangular ground truth regions, but in fact our system is able to capture most of the areas the pathologist focused on.

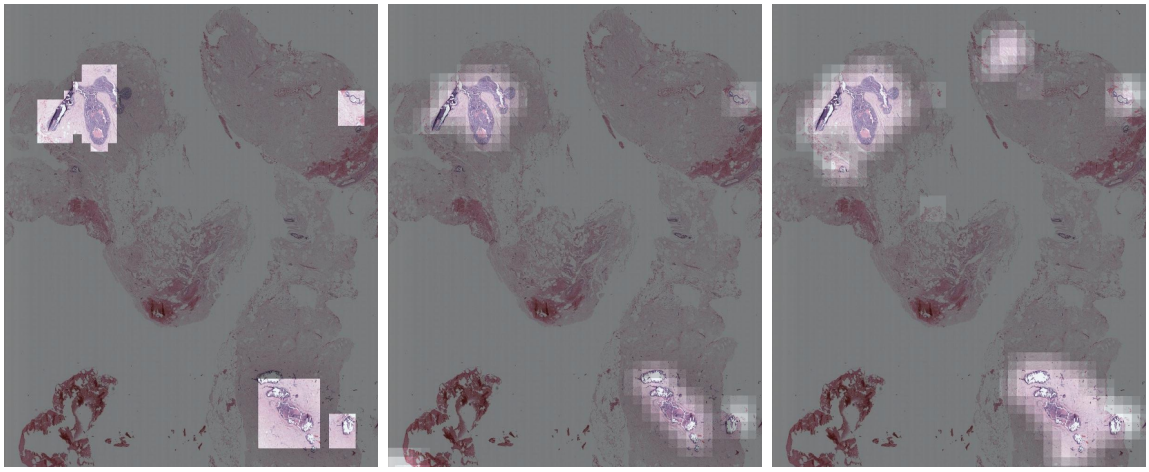


(a) Using ROIs extracted from viewport logs



(b) Using ROIs marked manually

Figure 4.7: ROI detection accuracies with different-sized visual dictionaries and different representations of visual words. The accuracies obtained by tenfold cross-validation experiments using **(a)** ROIs extracted through viewport analysis of three expert pathologists on 240 digital slides, **(b)** manually marked ROIs as training and viewport-extracted ROIs as test data. ©2016, Journal of Digital Imaging



(a) Ground Truth

(b) Predictions using ROIs extracted from viewport logs

(c) Predictions using ROIs marked manually

Figure 4.8: **ROI Detection Results (a)** Ground truth calculated by analyzing the viewport logs for a case. **(b)** Probability map showing the predictions made by using manually marked ROIs as training data and image patches as visual words. **(c)** Probability map showing the predictions made by using viewport extracted ROIs as training data and image patches as visual words. ©2016, Journal of Digital Imaging

4.5 Discussion

Whole slide digital images provide researchers with an unparalleled opportunity to study the diagnostic process of pathologists. Extracting regions of interest from whole slide images reduces the automated diagnosis problem into classification of ROIs. It also enables us to study the visual characteristics of the regions that attracted pathologists' attention without collecting manually labeled data.

Manually marked ROIs and viewport-extracted ROIs in our dataset have different semantics, the former denotes carefully selected areas displaying diagnostic characteristics of the WSI while the latter includes all ROIs that are taken into consideration during diagnosis. In other words, viewport-extracted ROIs are the candidates for the final manually marked ROIs. Yet, for the purposes of automated diagnosis or behavior analysis, viewport-extracted ROIs provide more insight.

In additional experiments, we analyzed the model with different-sized visual vocabularies. The dictionary size does not have an impact on the accuracy as long as the dictionary is large enough to

include basic building blocks of tissue images. Since breast histopathology images have less variability in comparison to every-day images, the dictionary size needed for a high detection accuracy is around 40 words — much smaller than general computer vision practices for the bag-of-words model. We also discovered that the words representing the epithelial cells are the most important words in representation of ROIs. When the dictionary size is decreased to 30-words where hierarchical clustering merges all epithelial cell clusters to others, the accuracy drops significantly. This is very intuitive since the breast cancer presents diagnostic features especially around epithelial structures of the tissue such as breast ducts and lobules.

We also experimented with a different visual word definition: superpixels. Using superpixels instead of square patches does not increase classification accuracy significantly. Furthermore, superpixels are computationally very expensive and slow in comparison to simple square patches.

A factor in our evaluation that should be considered is the nature of viewport-extracted ROIs. Because the tracking software records the portions of the digital slide visible on the screen, the viewports are always rectangular. Although this simple data collection allowed us to obtain a large dataset that is unique in the field, it has its shortcomings. In lower resolutions that correspond to small zoom levels, the viewports include a lot of surrounding uninteresting tissue (like background white space or tissue stroma) but there is no way to understand, outside of eye tracking, where the pathologist actually focused in these rectangular image regions. Our predictions, on the other hand, can be quite precise in ROI shapes.

Extracting ROIs from WSIs can serve as a filtering step that reduces the image size for more powerful yet computationally expensive image features. By designing image features specific to breast cancer diagnosis, it is possible to classify the ROIs detected, and thus diagnose the WSI. Our experiments showed that it is possible to detect ROIs using simple color and texture features and computationally expensive modifications do not improve the accuracy significantly.

4.6 Summary

Whole slide digital imaging technology enables researchers to study pathologists interpretive behavior as they view digital slides and gain new understanding of the diagnostic medical decision-

making process. In this study, we propose a simple yet important analysis to extract diagnostically relevant regions of interest (ROIs) from tracking records using only pathologists actions as they viewed biopsy specimens in the whole slide digital imaging format (zooming, panning, and fixating). We use these extracted regions in a visual bag-of-words model based on color and texture features to predict diagnostically relevant ROIs on whole slide images. Using a logistic regression classifier in a cross-validation setting on 240 digital breast biopsy slides and viewport tracking logs of three expert pathologists, we produce probability maps that show 74% overlap with the actual regions at which pathologists looked. We compare different bag-of-words models by changing dictionary size, visual word definition (patches vs. superpixels), and training data (automatically extracted ROIs vs. manually marked ROIs). This study is a first step in understanding the scanning behaviors of pathologists and the underlying reasons for diagnostic errors.

Chapter 5

TISSUE TYPE SEGMENTATION OF BREAST HISTOPATHOLOGY IMAGES

5.1 Introduction

5.1.1 Motivation

Segmenting whole slide images of breast biopsies into building blocks is crucial to understand the structural changes that lead to a diagnosis of breast cancer. Segmentation masks can provide information about the distribution and arrangement of different tissue types. Segmentation is a powerful abstraction for the detection and description of the structures.

5.1.2 Related Literature

Superpixels are frequently used in computer vision for segmentation of objects in images; Beck *et al.* conducted one of the first studies using superpixels in histopathological image analysis [12]. Using invasive breast cancer images and a large number of features extracted from superpixels, they discovered that stromal features are significant in survival of the patients. Their methodology depends on supervised classification of superpixels into several classes, including epithelial and stromal tissue, using labels from experts. Recently, Gorelick *et al.* used superpixel in prostate cancer detection and classification [39]. Similar to Beck *et al.*, Gorelick *et al.* also classified superpixels into semantically meaningful classes using expert labels. Superpixels were also adopted by some recent studies on localization of tumor in cancer images [4, 13].

Following the success of Convolutional Neural Nets (CNNs) in image classification tasks [44, 105, 118], they have been extended for high-level tasks such as semantic segmentation and activity recognition [66, 81, 86]. Medical image datasets are very small in size in comparison to natural image datasets. For instance, the gland segmentation dataset [102] has ≈ 90 images for training,

while the PASCAL VOC dataset [35] has $\approx$ 12000 images for training. Therefore, training CNNs from scratch for medical image analysis is not feasible. However, some work has applied CNNs for medical image analysis using transfer learning [21, 36, 92, 101]. Most notable among them are: classifying WSI into tumor subtypes and grades [45], segmenting EM images [92], segmenting gland images [21], and segmenting brain images [36]. Hou *et al.* [45] applied a sliding-window approach to reduce the WSI size and combine predictions made on patches to classify WSIs. Their work exploits WSI characteristics, such as the heterogeneity of tissue in terms of tumor grades and subtypes. Ronneberger *et al.* [92], Chen *et al.* [21], and Fakhry *et al.* [36] followed an encoder-decoder network approach with skip-connections for segmenting medical images. They studied the use of CNNs on datasets with limited input image sizes (for instance, the maximum dimension of the input in the gland segmentation dataset [102] is 775×552) and similarly sized objects of interest.

5.2 Tissue Labels

A set of labels was developed to describe the structural changes in the tissue on whole slide breast biopsy images. Although some of the labels are not important in diagnostic interpretation, the labels cover all the pixels in the images. The following set of seven tissue labels were selected with the input of expert pathologists:

- *Epithelium:* The ducts and lobules in breast are made up of epithelial cells. In order to isolate and describe the structures, epithelium needs to be segmented. We marked malignant and epithelial tissue separately but the distinction between two labels is difficult to make without additional stains on H&E images.

  - *Benign epithelium:* In benign breast tissue, epithelial cells line up at the edge of the ducts and lobules. In "normal" tissue, there are only two layers of cells around the ducts, but in cases mapping to the atypia category, the cells proliferate, leading to structural abnormalities. Although the epithelium in "atypia" cases shows cytological differences from that of the benign cases, it is still considered as benign epithelium

for our segmentation purposes.

  - *Malignant epithelium:* In DCIS and invasive cases, the epithelial cells are bigger and irregular in shape. These malignant epithelial cells keep dividing and create breast carcinomas either *in situ* or invasive.

- *Stroma:* Stroma is the supportive tissue that fills between the ducts in the breast. It consists of connective tissue, mainly collagen, and the blood vessels. Collagen has a pink appearance in H%E slides. Breast stroma can "react" to the cancer of the ducts. There is research showing that stromal appearance might affect the survival time in breast cancer [12].

  - *Desmoplastic stroma:* Desmoplasia is the growth of connective tissue in response to the presence of the cancer. In the case of breast cancer, desmoplasia is a reaction to the carcinoma. The cancer cells cause the proliferation of the fibroblasts and secretion of collagen. The tissue is rich with fibroblasts, which appear as smaller cells embedded in the connective tissue.
  - *Normal stroma:* We used the label stroma as a blanket term to cover stroma tissue that is not desmoplastic.

- *Secretion:* The ducts and lobules are parts of a glands responsible for producing the milk and deliver it to the nipple. In the case of benign and atypia categories, the ducts and lobules might be filled with molecules discharged from the cells. We used secretion as a label to mark any substance filling the ducts.

- *Necrosis:* Necrosis is the death of the cells in the tissue. In the case of breast pathology, the cells in the carcinoma dies due to lack of blood supply caused by the abnormal growth of the tumor cells. Necrosis usually happens in DCIS cases, close to the center of the duct, which is filled with proliferating malignant cells.

- *Blood:* Although not diagnostically significant, it is common to see vessels and red blood cells in breast biopsy slides. These are very small regions but are very distinct in appearance.

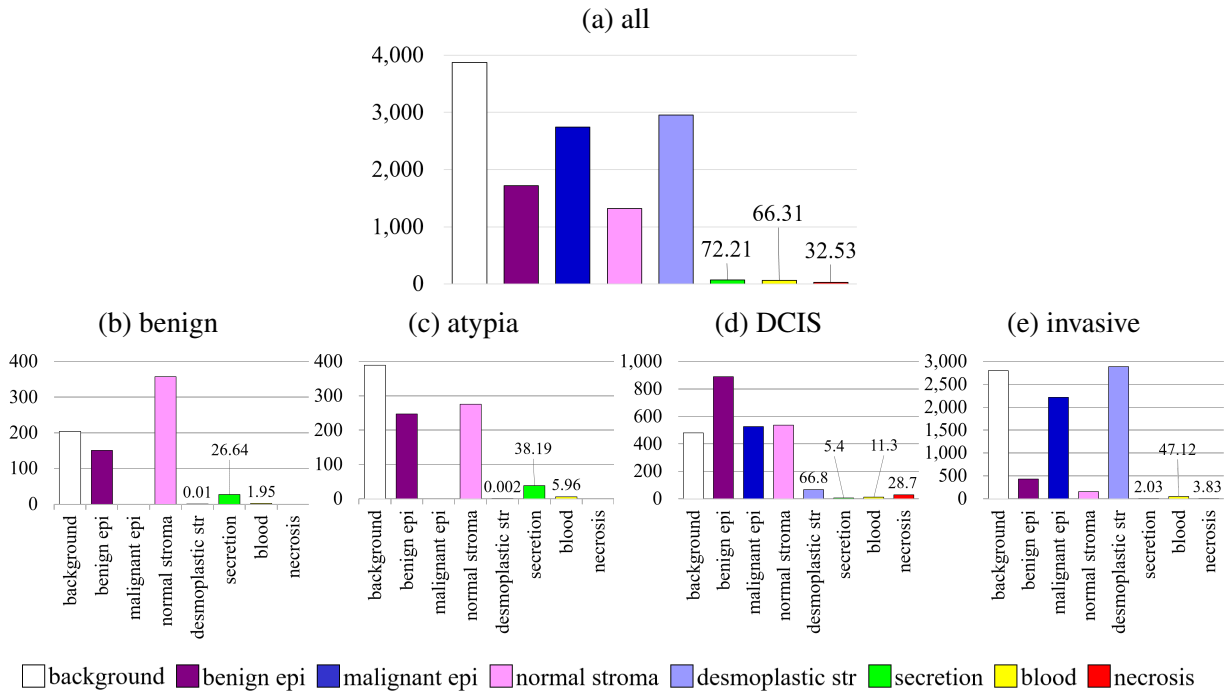


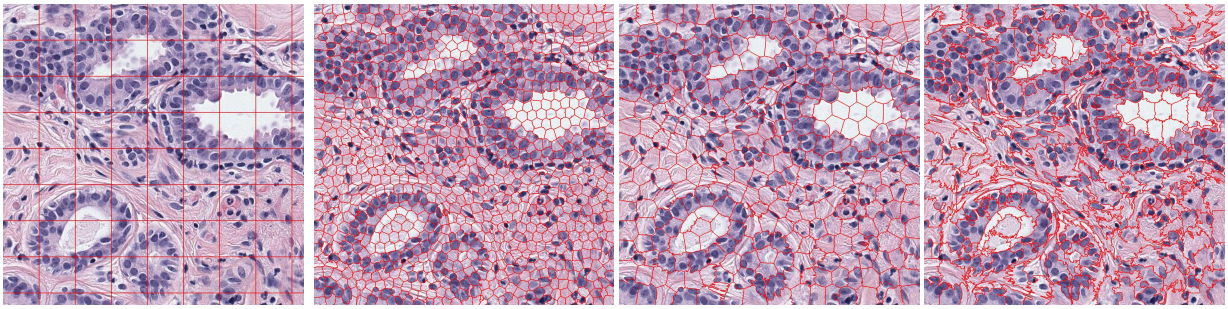
Figure 5.1: Distribution of tissue labels in all ROIs (a), and in individual diagnostic categories of benign (b), atypia (c), DCIS (d) and invasive (e).

Since labeling is a very expensive task considering the expertise needed, a subset of 40 of the 240 cases were selected to be labeled by a pathologist. We used the variables of diagnosis, breast density and age in the selection. 58 ROIs obtained from these 40 cases were labeled with seven tissue labels. Figure 5.1 shows the distribution of labels in the ROIs marked by the expert.

5.3 Superpixels and SVM for Segmentation

5.3.1 Superpixel Segmentation

In Chapter 4, we compared 120x120 pixel image patches and superpixels as *visual words* in a bag-of-words model for ROI detection task. Our analysis showed that using superpixels does not improve the accuracy of ROI detection significantly. Although image patches are an easy and efficient way of building visual dictionaries, they can divide biological structures (e.g. cells,

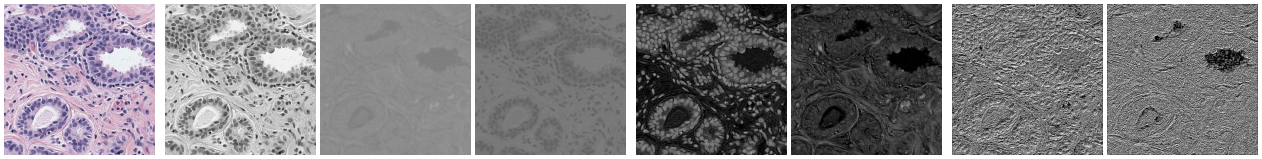


(a) (b) (c) (d)

Figure 5.2: Superpixels vs patches: (a) Image patches divide an image into visual words without using content information. The SLIC [3] algorithm can find superpixels with different sizes and compactness. (b) shows a denser segmentation than (c), (d) shows a less compact segmentation than (c).

lumen, ducts). In this chapter, we aim to segment the biopsy images into different "tissue types". Superpixels are better suited for this task, because they can take any shape and their borders align with the structures in the tissue. See Figure 5.2 for a comparison of image patches to superpixels.

We used the efficient implementation of simple linear iterative clustering (SLIC) [3] for segmenting RGB images into superpixels. SLIC has parameters controlling the size and the compactness of the superpixels. According to the SLIC algorithm, there are two measures minimized for the pixels in a superpixel: 1) the proximity on *xy* plane and 2) the proximity in *color space*. The compactness parameter adjust the trade-off between color similarity of the pixels and the shape of the superpixel. If compactness parameter is high, the pixels in a superpixel are forced to be closer at the expense of the color similarity constraint. A low compactness parameter causes superpixels to have any shape as long as the colors of the pixels in the same superpixel are similar. Figure 5.2 shows example outputs with different parameters. For the purposes of this work, we used a set of parameters that give us a superpixel segmentation similar to Figure 5.2b. One factor in parameter selection was that a superpixel should contain at least one epithelial cell and we achieved this by using superpixels of size 3000s pixel. This configuration decreases dimensionality of the image by approximately 3000 pixels to one superpixel. We calculated color histograms in L*a*b* and H&E color spaces and LBP texture histograms for each superpixel (Figure 5.3).



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 5.3: Image features calculated from superpixels: (a) RGB image, (b,c,d) L*a*b* channels, (e-f) H&E channels, (g-h) LBP feature of H&E channels.

5.3.2 Circular Neighborhoods

By using superpixels, we are defining the semantic segmentation task as the classification of small regions with an approximate area of 3000-pixel. This simple approach has been used by Beck *et al.* to discover image features predicting the survival in breast cancer [12]; however, for their purpose of high-level analysis, an inaccurate segmentation is acceptable. By visualizing the results of our preliminary experiments, we found that some of the error was due to the mis-prediction of individual superpixels surrounded by correctly predicted superpixels. This was due to the fact that the information available to the classifier was limited to the superpixels. Humans, on the other hand, use context information when assigning the tissue labels to the pixels.

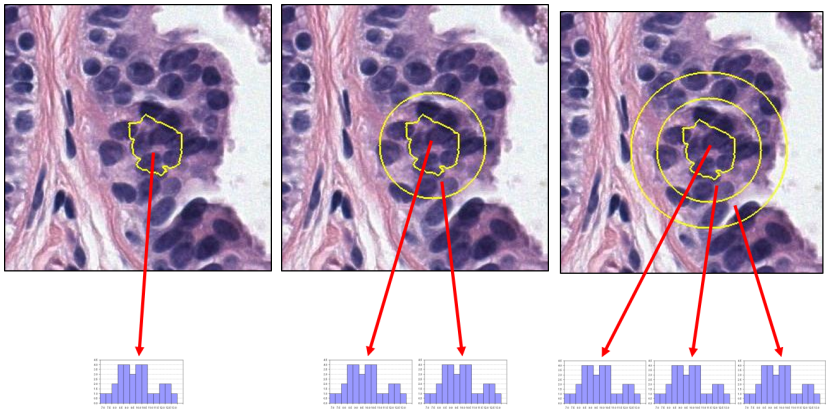
We improved our superpixel classification results by including features extracted from the surrounding tissue. We extracted features from two circular neighborhoods centered at the superpixel's center. By using two circles, we aimed to let the classifier learn the importance of the immediate neighbors and the farther neighbors.

5.3.3 Experiments and Results

Experimental Setup

We trained Support Vector Machine (SVM) classifiers [25] for three feature configurations:

1. *Superpixels only*: Color and texture histograms extracted from only the superpixel are used. (Figure 5.4a)



(a) Superpixel only

(b) Superpixel + 1 neighborhood

(c) Superpixel + 2 neighborhoods

Figure 5.4: Circular neighborhoods of a superpixel included in feature extraction to improve the classification.

2. *Superpixel + 1 neighborhood*: Two color and texture histograms extracted from the immediate circular neighborhood and from the superpixel are concatenated. (Figure 5.4b)

3. *Superpixel + 2 neighborhoods*: Three color and texture histograms extracted from the immediate circular neighborhood, from the second circular neighborhood and from the superpixel are concatenated. (Figure 5.4c)

The complete dataset contains about 3.8 million superpixels, each corresponding to a sample with a color and texture feature vector. Because of the non-uniform distribution of the tissue labels and ROI size variation, we sampled 2000 superpixels for each of the eight labels (if possible) from each of the 58 ROIs. Subsampling ensured the classifier had similar number of samples from each class and from each image. We trained SVM classifiers in a 10-fold-cross-validation setting by including similar numbers of samples from each image in each fold to ensure that the classifier saw samples from each image in each fold.

Evaluation Metrics

Precision, also called positive predictive value, is the fraction of instances correctly classified as a label among all the instances classified as belonging to that class. *Recall*, also called sensitivity, is the fraction of instances correctly classified as a label among all the instances actually belonging to that class. *Jaccard Index*, also known as *Intersection over Union*, is a summary statistics that is based on both false positives and false negatives, while the precision and recall consider only one.

Precision, recall and Jaccard Index are complementary statistics measuring different aspects of the classification performance. Below is a summary of the three statistics:

$$\left| \begin{array}{c} \text{Precision} = \frac{tp}{tp+fp} \\ = \frac{\text{\# pixels correctly classfified as label A}}{\text{\# pixels classified as label A}} \end{array} \right| \left\| \begin{array}{c} \text{Recall} = \frac{tp}{tp+fn} \\ = \frac{\text{\# pixels correctly classfified as label A}}{\text{\# pixels in label A}} \end{array} \right\| \left| \begin{array}{c} \text{Jaccard Index} = \frac{tp}{tp+fp+fn} \\ = \frac{\text{\# pixels correctly classfified as label A}}{\text{\# pixels classified as or actually in label A}} \end{array} \right|$$

tp = true positives, fp = false positives

tn = true negatives fn = false negatives

F_1 score, also known as F-measure is another measure of accuracy that combines both the precision and recall. F_1 score is the harmonic mean of precision and recall – multiplying the constant of 2 scales the score to 1 when both recall and precision are 1.

$$F_1 \text{ score} = 2\times \frac{1}{\frac{1}{\text{precision}}+\frac{1}{\text{recall}}} = 2\times \frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$$

Results

Table 5.1 shows the average class Jaccard Index and F_1 score for the three feature combinations. Both metrics increase when more context is included in feature extraction.

Figure 5.5 shows the confusion matrices for the three models. We can make following observations for all models:

- The background was segmented very accurately. The majority of the mis-prediction for the background pixels were for the secretion label. This may be due to the labeling error (when

Table 5.1: Average class segmentation accuracies for three methods using different superpixel neighborhoods to extract features.

Context included	*Avg. class accuracy*	
	Jaccard Index	*F_1 score*
Superpixel only	.22	.33
Superpixel + 1 neighborhood	.23	.34
Superpixel + 2 neighborhoods	.24	.35

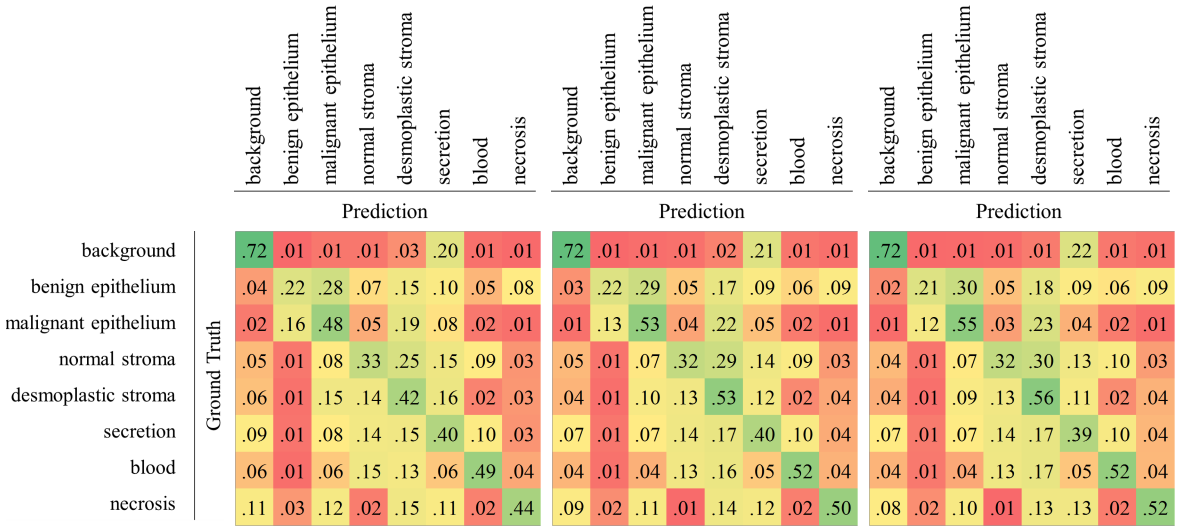
pathologist marked a region as secretion, disregarding small background regions inside or around the secretion) or similarity in the appearance (color and texture) of background and secretion.

- Benign epithelium was confused with malignant epithelium. Although the texture feature can pick up the differences between benign and malignant cells, in clinical practice, the differentiation is usually made at a cellular level using features like cell size, shape and circularity. Simple color and texture features may be insufficient for this differentiation.

- The two stroma labels, normal stroma and desmoplastic stroma, were confused with each other. This maybe due to the labeling error or the similar appearance of the two labels.

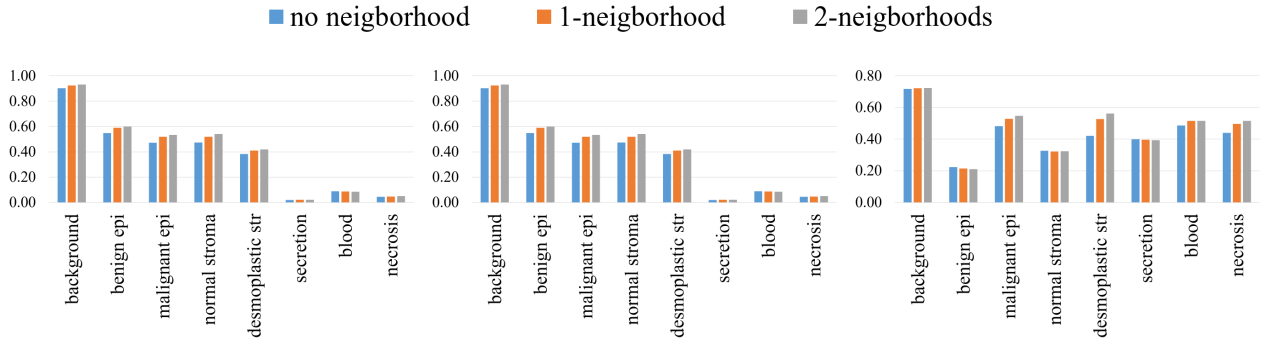
- Desmoplastic stroma was also confused with malignant epithelium. Desmoplastic stroma is usually filled with fibroblasts (the cells that produce the connective tissue fibers) that appear as small dark purple round shapes. They are also dyed by the haematoxylin that gives the characteristic blue color to all cells in H&E slides. This color is primarily dominant in epithelium labels.

- Secretion, blood and necrosis labels were confused with the background and two stroma labels. Since these three are the rarest labels and usually surrounded by either stroma or background, this maybe due to labeling error (when pathologist missed some of the secretion and it was counted towards background in the ground truth labels) or inadequate number of positive samples.



(a) superpixel only (b) superpixel + 1 circle (c) superpixel + 2 circles

Figure 5.5: Confusion matrices for three superpixel-based models using color and texture histograms extracted from (a) only the superpixel, (b) the superpixel and the first circular neighborhood, and (c) the superpixel and two circular neighborhoods. The rows are ground truth labels and the columns are predictions. The percentages of the all pixels in a ground truth label is reported, i.e. diagonal values give the recall rates for each label.



(a) Jaccard Index (b) Precision (c) Recall

Figure 5.6: Jaccard Index, precision and recall rates for the three superpixel-based models using color and texture histograms extracted from (a) only the superpixel, (b) the superpixel and the first circular neighborhood, and (c) the superpixel and two circular neighborhoods. All metrics increase when more context is included in the segmentation.

Figure 5.6 shows the Jaccard Index, precision and recall metrics for the eight labels with the different models. The recall is high for all labels whereas precision is low for secretion, blood and necrosis labels. High recall for a label indicates that most of pixels with that label in the ground truth are detected in the prediction, high precision for a label means that from the pixels predicted as that label, most of them have that label in the ground truth. In other words, low precision for secretion, blood and necrosis indicate a lot of false positives for these labels. Jaccard Index is low for the secretion, blood and necrosis classes because there aren't many positive samples for these classes in the dataset to begin with. Thus any false positive affects the precision and Jaccard Index more than other classes with many samples.

Figure 5.7 shows example visualizations from the three model evaluated. The false positives, especially for the secretion and necrosis labels, are apparent in the predictions. Including context by adding circular neighborhoods into feature vectors improves the prediction and acts as a noise reduction technique.

Another detail to note is the ragged borders caused by the superpixel segmentation. Since our evaluations are pixel-based and the ground truth labels have smooth borders, these ragged borders introduce some error independent of the classification. This is a shortcoming of the superpixel segmentation strategy that we use as a first step. The next section, Section 5.4 introduces a deep-learning methodology that uses raw RGB pixel values instead of color and texture features of superpixels.

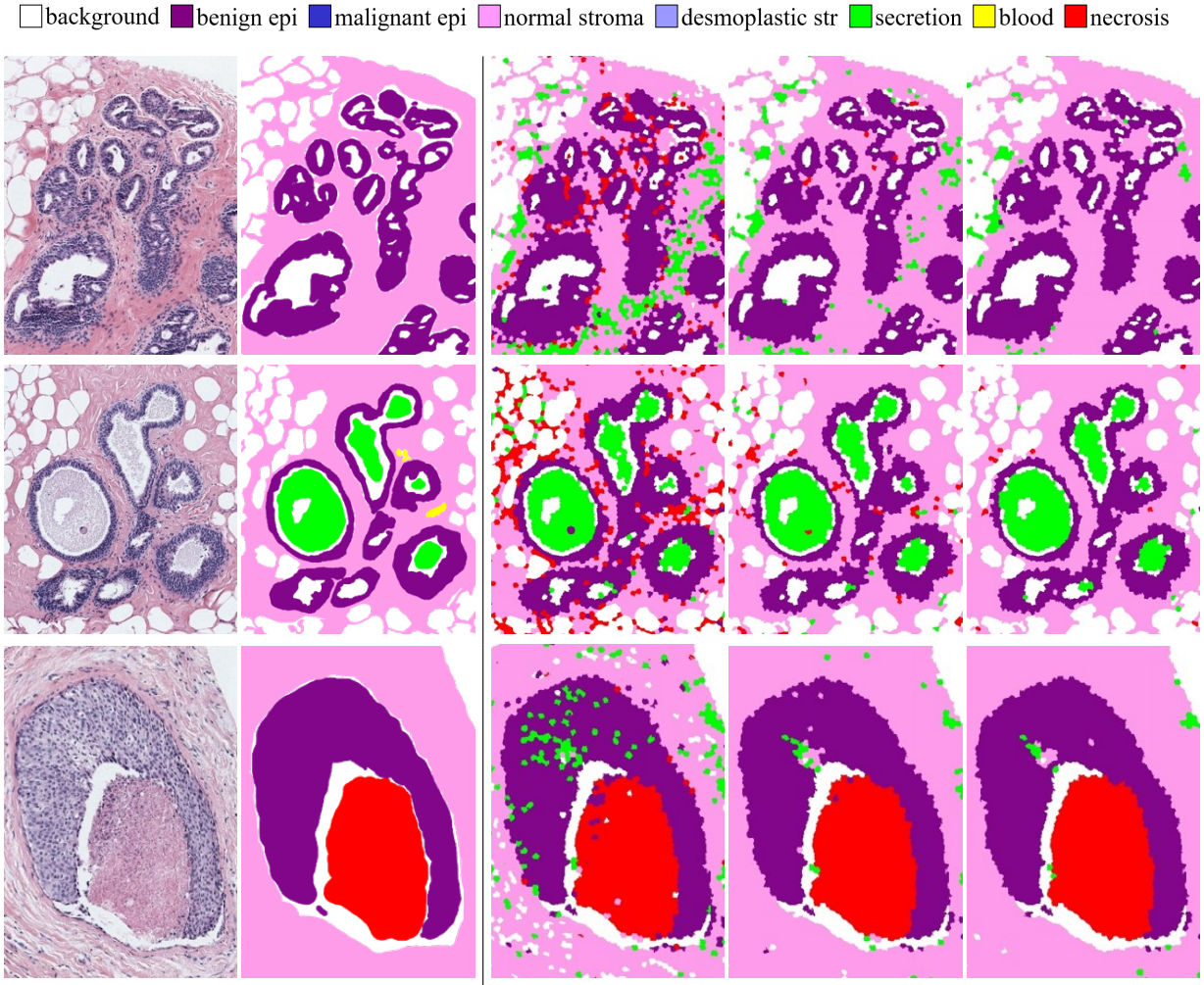


Figure 5.7: Visualizations of the segmentations produced by the three superpixel-based methods: (a) input image, (b) ground truth labels, (c) the prediction made by using only the superpixel region, (d) the prediction made by using the superpixel region and the first circular neighborhood, (e) the prediction made by using the superpixel and the two circular neighborhoods. Using more context around the superpixel improves the accuracy and corrects mis-predicted small regions.

5.4 Convolutional Neural Nets for Segmentation

Convolutional neural networks (CNNs), the most popular choice for classifying, detecting, and segmenting natural images [44, 66, 81, 105, 118], often lack the large scale, labeled datasets needed

for medical image analysis. Considerable effort has been applied toward producing public datasets for medical tasks such as TCGA [2], mitosis detection [95, 112], cell segmentation [7], and gland segmentation [102]. Such datasets have fixed-sized input images and, more importantly, similar sized objects of interest; this allows the successful application of CNNs for medical image analysis [21, 36, 92].

Unlike existing datasets, the digiPATH dataset has images and objects in the images that are variable in size due to wide range of diagnostic categories. Therefore, segmentation of WSIs have particular challenges:

- The limited computational resources make learning directly from large, variably sized images using CNNs difficult. Reading a digital WSI into memory, let alone a batch of training samples, is impossible with commodity hardware. Previous applications of CNNs for segmentation of medical images used hand-cropped regions in reasonable sizes.

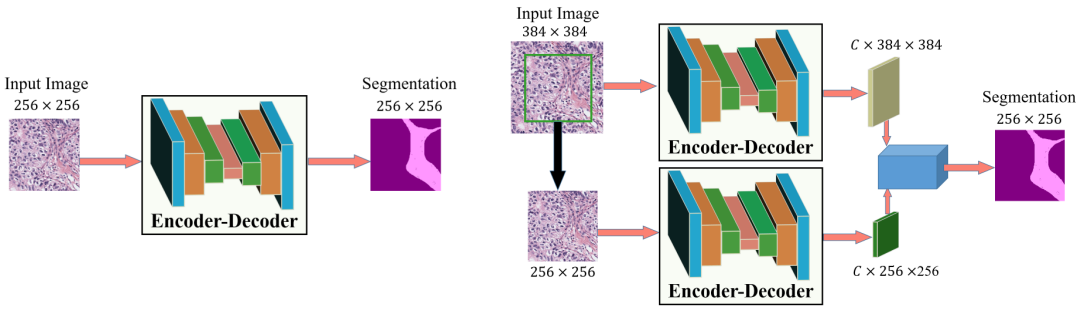
- A sliding-window approach is promising for segmenting WSIs; however, the size of the patch determines the context available to the model. This approach divides the bigger structures into smaller patches and some structures may cover several patches. This may hurt the performance of the segmentation method. Furthermore, objects or biological structures come in different sizes, and simple scaling strategies do not work.

- The rarity of some biologically important labels makes them hard to detect. Less than 1% of all pixels are necrosis and secretion, yet these labels may be important in diagnostic process.

We developed three CNN-based segmentation models:

Plain Network

We implemented a simple encoder-decoder architecture with residual connections following the work of Fekhry *et al.*[36], illustrated in Figure 5.8a. The details of the encoder-decoder are given in Appendix B.2, and it is illustrated in Figure B.2a.



(a) Plain-Network (b) Multi-Resolution-Network

Figure 5.8: Two network architectures: (a) Plain-Network, a simple encoder-decoder network, and (b) Multi-Resolution Network that combines multiple resolutions of the input image.

MR-Plain Network

We implemented a multi-resolution network by using the Plain Network on two different resolutions of the input image and aggregating the results with a short network, illustrated Figure 5.8b. One network takes a larger image than the other one and the result of the larger image is cropped to match the other one at the end. Figure 5.9 compares the inputs of the two encoder-decoders. Note that the aggregation of the two segmentations is also learned.

MR-MP-Network

The final CNN-based model we implemented introduces a novel encoder-decoder that is described in detail (Appendix B.4) and illustrated in Figure B.2b. We used the multi-resolution network from Figure 5.8b but replaced the encoder-decoder with an architecture that used multiple paths and shared the information between different layers of encoders with input-aware encoding blocks.

5.4.1 Experiments and Results with Five Labels

We trained and compared four methods:

- Superpixel-based SVM: Explained in Section 5.3.

Table 5.2: Distribution of diagnostic categories and ROI size variance for different diagnoses

Diagnostic Category	*#ROIs* (total)	*#ROIs* (train)	*#ROIs* (test)	*Avg. ROI size*
benign	9	4	5	$9K \times 9K$
atypia	22	11	11	$6K \times 7K$
DCIS	22	12	10	$8K \times 10K$
invasive	5	3	2	$38K \times 44K$
Total	58	30	28	$10K \times 12K$

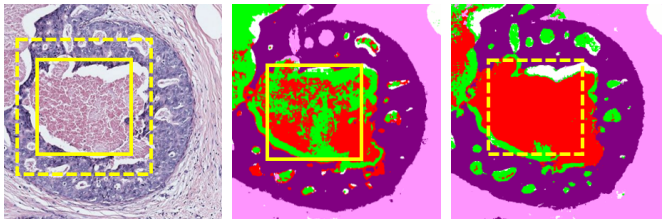
- Plain-Network: A plain network (Figure 5.8a) with a simple encoder-decoder architecture (Figure B.2a).
- MR-Plain-Network: A multi-resolution network (Figure 5.8b) with two simple encoder-decoders (Figure B.2a).
- MR-MP-Network: A multi-resolution network (Figure 5.8b) with two multi-path encoder-decoders using input-aware encoding blocks (Figure B.2b).

Training Details

We split 58 ROIs into training (30 ROIs) and test (28 ROIs) sets keeping the distribution of diagnostic categories similar (Table 5.2).

For the plain encoder-decoder network, we cropped patches of size 256×256 with an overlap of 56 pixels. For the multi-resolution network, for each 256×256 patch, we created another patch by including a 64-pixel border area (see Figure 5.9). When necessary, we used symmetric padding to complete the patches. We obtained $5,312$ patches from 30 ROIs. We used random rotations (between 5 and 10 degree), horizontal flips, and random crops followed by scaling (i.e. the crop border is selected randomly between 20 and 50 pixels) to augment the data, resulting in a total of 25,992 patches that were split into the training and validation sets using 90:10 split ratio.

We used five training labels: background, epithelium, stroma, secretion and necrosis. For simplicity, we merged two epithelium and two stroma classes, and added blood cells to the stroma class.



(a) (b) (c)

Figure 5.9: Context available to the Plain and Multi-Resolution Networks. (a) The patches used by the networks, the patch with solid borders is used by the Plain-Network while the patch with the dashed border is included in the Multi-Resolution Network (MR-Plain-Network and MR-MP-Network). (b) The segmentation produced by using the patch with solid borders in (a). (c) The segmentation produced by using the combination of solid and dashed bordered patches in (a). Note that the segmented patch is the same size in (b) and (c).

We trained all of our models end-to-end (until they converged) using stochastic gradient descent with a fixed learning rate of 0.0001, momentum of 0.9, weight decay of 0.0005, and a batch size of 10 on a single NVIDIA GTX-1080 GPU. We initialized encoder weight with ResNet-18 [118] trained on the ImageNet dataset. We choose ResNet-18 because it: (1) is fast at inference, (2) requires less memory per image, and (3) learns less parameters while delivering accuracy similar to VGG-16 [44] on the ImageNet. We initialized decoder weights as in [43]. We applied batch normalization [48] after every convolutional and deconvolutional layer to reduce the internal-covariate-shift. We did not use dropout, following the practice of [118]. We evaluated our methods on the test set using standard evaluation metrics such as precision, recall, F-score and Jaccard Index[1] [21]. All of our models were trained end-to-end using the Torch framework.

Results

We evaluated three networks and superpixel-based method from Section 5.3 on the 28 test ROIs. We calculated four metrics from the prediction results: precision, recall, F_1 score and Jaccard Index (see Section 5.3.3 for details).

[1]We used the Jaccard index (J_I) instead of the Dice index (D_I) that was used in [21]. Since the Jaccard and Dice indices are monotonic, we can convert from one to the other easily. $D_I = \frac{2 \times J_I}{J_I + 1}$

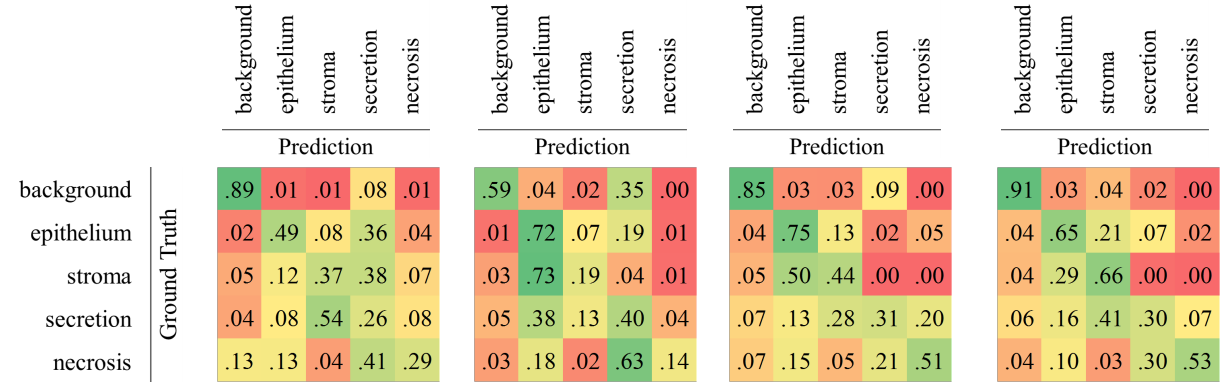
Table 5.3 shows average Jaccard Index and F_1 scores for the four methods we compared: Superpixels + SVM, Plain-Network, MR-Plain-Network and MR-MP-Network. The average Jaccard Index and F_1 score decreased from Superpixels + SVM to Plain-Network. It maybe due to fact that the sample size required to train a CNN is much larger than an SVM and Plain-Network is not a powerful enough model to generalize the RGB features to labels whereas SVM uses more powerful hand-crafted features of color and texture histograms with the superpixel segmentation. MR-Plain-Network outperformed Superpixels + SVM thanks to the multi-resolution network architecture even though the encoder-decoder was the same as Plain-Network. By increasing the context available to the CNN with multi-resolution network, we increased the average Jaccard Index by .11 and the average F_1 score by .06. The best performing model was the MR-MP-Network with the multi-resolution architecture, multi-path encoder-decoder and input-aware encoding blocks. By using the proposed encoder-decoder instead of the plain one, we improved the average Jaccard Index of MR-Plain-Network by .03 and F_1 score by .05.

Table 5.3 also shows the number of parameters and inference time for the CNN-based methods. As expected, the improvement in the average class accuracy costs time during inference. By adding more blocks and convolutions, we increased the number of parameters from 12.88 million in Plain-Network to 25.6 million in MR-Plain-Network and 25.76 million in MR-MP-Network.

Table 5.3: Comparison of the superpixel-based and three CNN-based models in terms of number of parameters, inference time and average class accuracy for the segmentation with five labels.

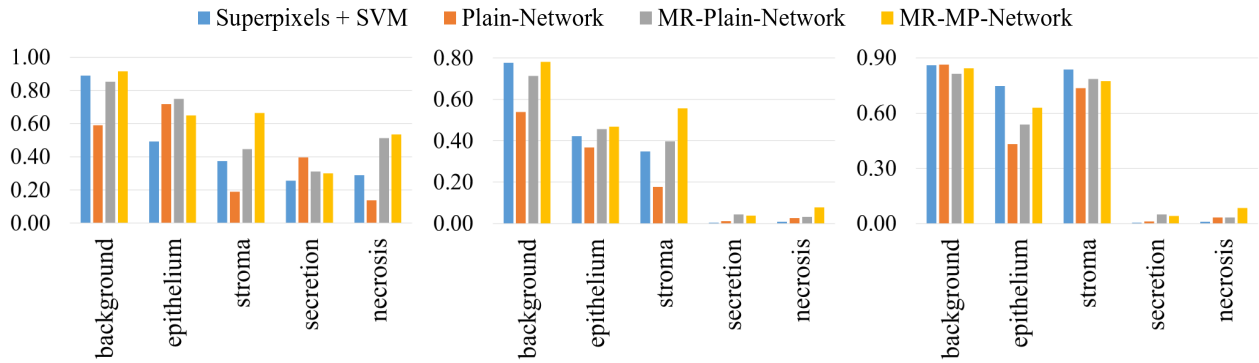
Segmentation Model	*# Parameters (in million)*	*Inference Time (in milliseconds)*	*Avg. class accuracy*	
			Jaccard Index	*F_1 score*
Superpixels +SVM	NA	NA	.20	.28
Plain-Network	**12.88**	**56.68**	.15	.23
MR-Plain-Network	25.6	64.99	.23	.34
MR-MP-Network	25.76	74.15	**.26**	**.39**

Figure 5.10 shows the confusion matrices for the four methods evaluated on the test set. Note that the confusion matrix for the Superpixels + SVM method is different from the one in Section 5.3.3 because the test sets were different. To be able to compare the four methods, we tested



(a) Superpixels + SVM (b) Plain-Network (c) MR-Plain-Network (d) MR-MP-Network

Figure 5.10: Confusion matrices for the superpixel-based and three CNN-based models: (a) Superpixels + SVM model, (b) Plain-Network, (c) MR-Plain-Network and (d) MR-MP-Network. The rows are ground truth labels and the columns are predictions. The percentages of all the pixels in a ground truth label is reported, i.e. diagonal values give the recall rates for each label.



(a) Jaccard Index (b) Precision (c) Recall

Figure 5.11: Jaccard Index, precision and recall rates for the superpixel-based and three CNN-based models: (a) Superpixels + SVM model, (b) Plain-Network, (c) MR-Plain-Network and (d) MR-MP-Network.

the Superpixels + SVM method on the same test set as CNNs. The Superpixels + SVM method confused epithelium, stroma and necrosis labels with the secretion label. Especially in the case of stroma, most of the stroma pixels were predicted as secretion. On the other hand, the Plain-Network struggled with stroma label, predicting 72% of the stroma pixels as epithelium. This trend is also visible in the visualizations in Figure 5.12. Another difficult label for the Plain-

Network was necrosis; it predicted 63% of the necrosis pixels as secretion. From Plain-Network to the MR-Plain-Network, the segmentation of all the labels except secretion improved. Although improved from Plain-Network, the MR-Plain-Network still confused stroma with epithelium by predicting half of the stroma pixels as epithelium. Finally, the MR-MP-Network outperformed all the other models in the segmentation of background, stroma and necrosis. The MR-MP-Network outperformed Superpixels + SVM baseline in all labels. It did worse than other networks in the epithelium label with an apparent trend of predicting the epithelium, stroma and secretion pixels as stroma. This trend can be observed in the visualizations on the second row of Figure 5.12 where the epithelium pixels (purple) were predicted as stroma (pink) at the last column.

Figure 5.11 shows the Jaccard Index, precision and recall of the four methods compared for the five tissue labels. All method had low recall for the secretion and necrosis labels due to the small sample size for these labels. However, the MR-MP-Network outperformed the other three methods by a margin in the necrosis label. Similarly, for the stroma label, the MR-MP-Network had high precision indicating less false positives together with high recall indicating less false negatives.

Figure 5.12 demonstrates some of the challenges of using CNNs to segment WSIs. The results for the Plain-Network (fourth column), the naivest CNN-based approach, illustrates the stitching effect especially in the second row where a larger ROI is shown. Using the multi-resolution network in the MR-Plain-Network and MR-MP-Network, we overcame this stitching effect that causes sharp line borders where the patches meet.

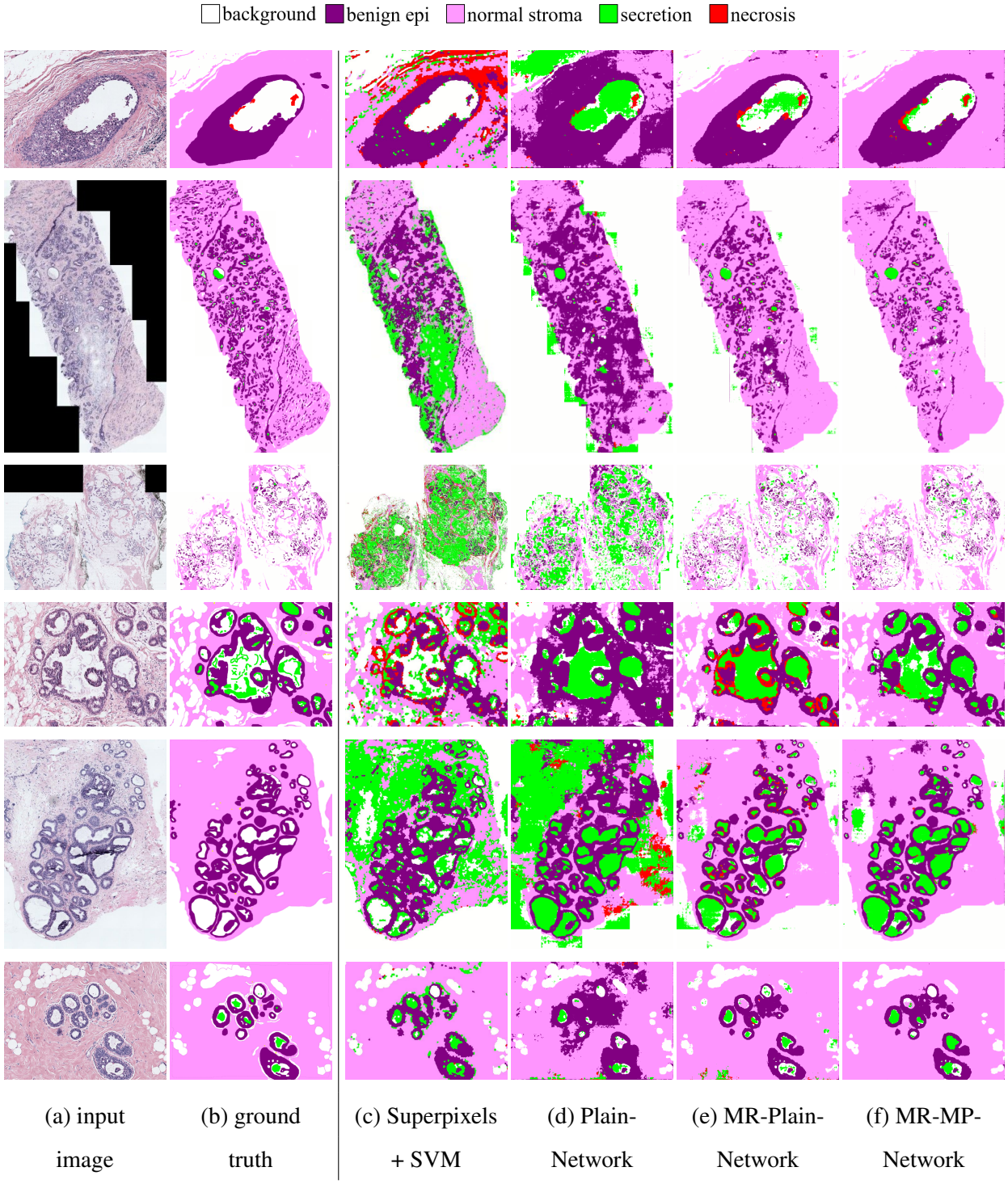


Figure 5.12: Visualizations of the segmentations produced by the superpixel-based and three CNN-based models:: (a) input image, (b) ground truth labels, (c) the prediction made by the Superpixels + SVM model, (d) the prediction made by the Plain-Network, (e) the prediction made by the MR-Plain-Network, and (f) the prediction made by the MR-MP-Network.

5.4.2 *Extension to Seven Label Classes*

Our ground truth labels include seven (eight including background) tissue classes; but we used only four (five including background) labels in the training and evaluation of the CNN-based methods to provide enough samples for each label during the training of the convolutional networks. However, the differentiation of benign and malignant epithelium, of normal and desmoplastic stroma may be still useful for automated diagnosis. By fine-tuning the best-performing CNN, MR-MP-Network, we aimed to segment the eight tissue labels.

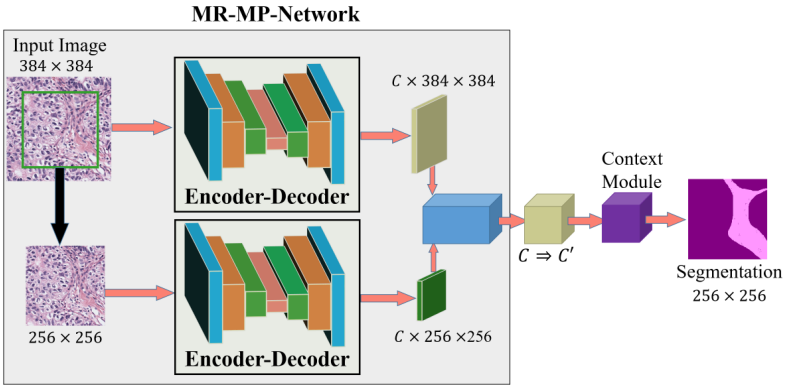


Figure 5.13: The extension of the MR-MP-Network to segment eight label classes and the context module developed using residual blocks with dilated convolutions.

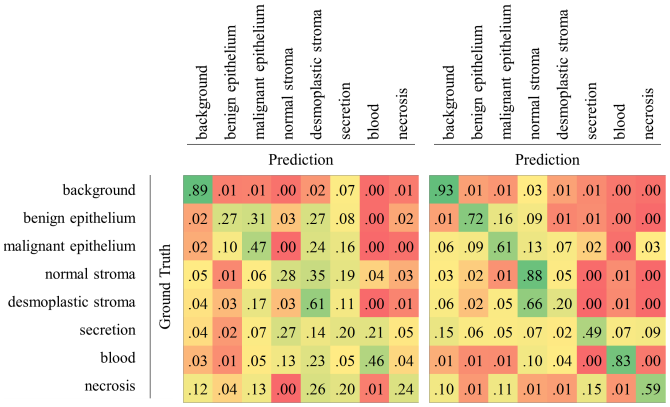
Starting with the pre-trained MR-MP-Network, we first projected the feature vector of the MR-MP-Network with C=5 labels to C'=8 labels. Then using a series of residual blocks with dilated convolutions, we built a 'context module' to improve the segmentation. The context module increased the receptive field of the neurons, improving the segmentation borders. Figure 5.13 shows the high-level architecture. For the encoder-decoder blocks, we used the multi-path encoder-decoder from Figure B.2b.

Experiments and Results with Seven Labels

Table 5.4 shows the average Jaccard Index and F_1 score for the two methods. CNN-based method achieves a better Jaccard Index and F_1 than superpixel-based method. Figure 5.14 shows the con-

Table 5.4: Comparison of the superpixel-based and three CNN-based models in average class accuracies for eight labels.

Segmentation Model	*Average Jaccard Index*	*Average F_1 score*
Superpixels +SVM	.29	.40
MR-MP-Network	.37	.50



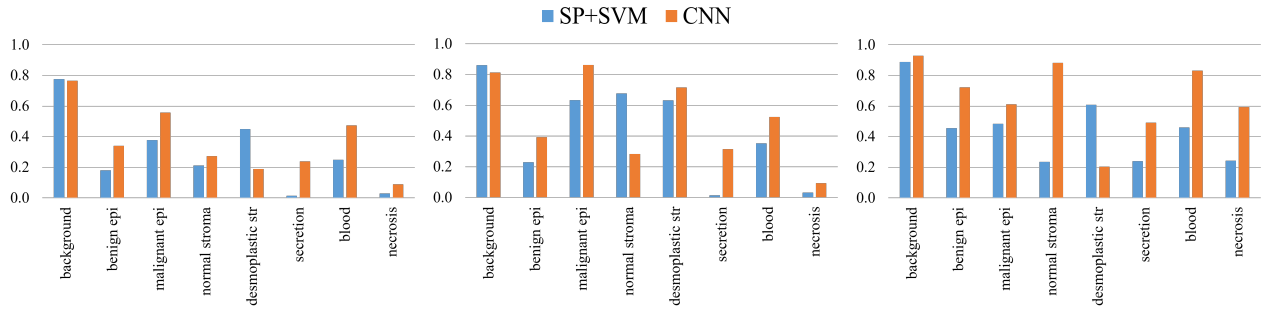
(a) Superpixels + SVM

(b) MR-MP-Network

Figure 5.14: Confustion matrices for the superpixel-based and the CNN-based model: (a) Superpixels + SVM model, (b) MR-MP-Network. The rows are ground truth labels and the columns are predictions. The percentages of the all pixels in a ground truth label is reported, i.e. diagonal values give the recall rates for each label.

fusion matrices for the two methods. CNN-based MR-MP-Network methods does better than the Superpixel + SVM method in every label, other than the desmoplastic stroma label. The MR-MP-Network confuses the desmoplastic stroma label with the normal stroma label. Both methods confuse two epithelium labels, benign and malignant epithelium, with each other.

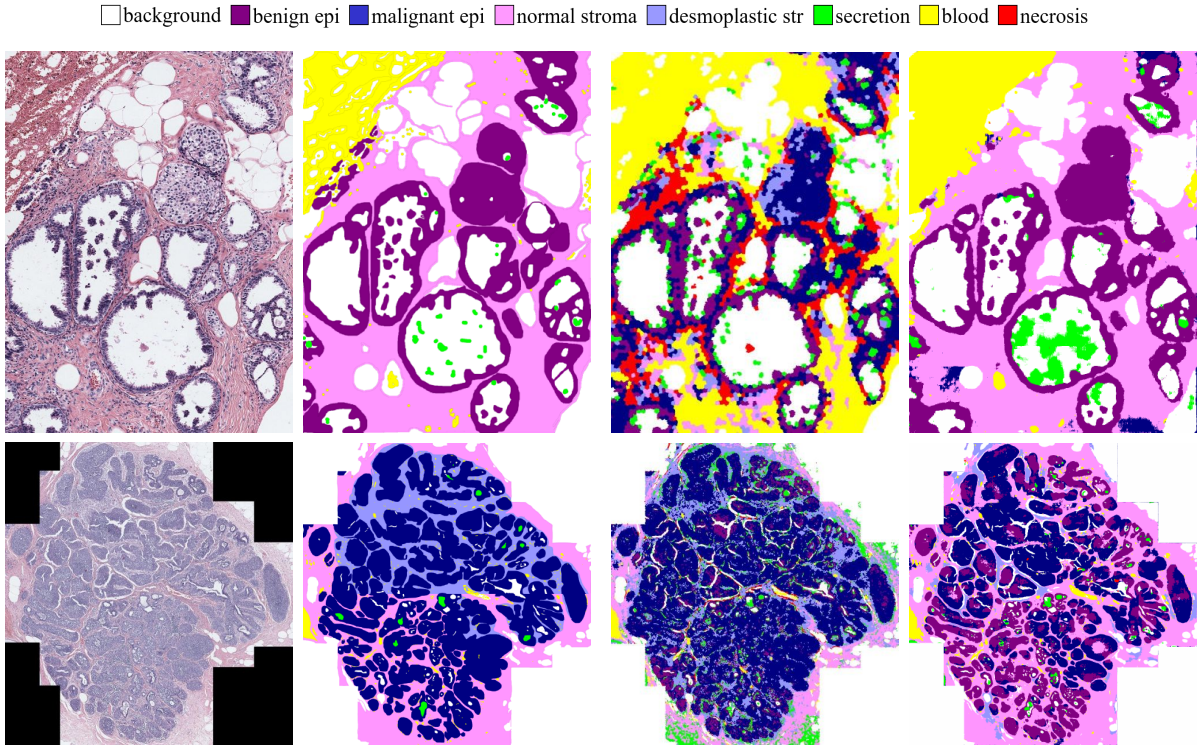
The CNN-based method performs especially well with the rare labels of secretion, blood and necrosis. Figure 5.15 compares the Jaccard Index, precision and recall statistics of the twp methods for each class. The CNN-based method has a high precision and recall in the secretion, blood and necrosis labels. However, it suffers from a low precision, high recall of the normal stroma label. This may be due to the labeling the majority of the desmoplastic stroma pixels as normal stroma.



(a) Jaccard Index (b) Precision (c) Recall

Figure 5.15: Jaccard Index, precision and recall rates for the superpixel-based and three CNN-based models: (a) Superpixels + SVM model, (b) Plain-Network, (c) MR-Plain-Network and (d) MR-MP-Network.

Finally, Figure 5.16 shows two example results from the two methods. The first row shows an example where the CNN-based segmentation model outperforms the superpixel-based segmentation model. The second row shows an example where the CNN-based model fails to classify desmoplastic stroma pixels correctly. Note that desmoplastic stroma occurs around tumors and the invasive cancer images in our dataset are much larger than other images. These evaluations are based on the small labeled subset (N=58) of the full digiPATH data (N=428) we have. Even though the CNN-based method performed poorly on the desmoplastic stroma label on this development set, the qualitative analysis of the results and the average scores show the strength of the CNN-based method.



(a) Input Image (b) Ground Truth (c) Superpixels + SVM (d) MR-MP-Network

Figure 5.16: Visualizations of the segmentations produced by the superpixel-based and CNN-based models using the eight tissue labels: (a) input image, (b) ground truth labels, (c) the prediction made by the Superpixels + SVM model, (d) the prediction made by the MR-MP-Network.

5.5 Discussion

SVM vs CNN: Impact of Context

The CNNs outperformed many of the traditional models composed of a classifier and hand-crafted image features in classification, detection and segmentation tasks. Our experiments showed that it is possible to obtain a performance boost by using an encoder-decoder architecture specifically designed for the breast biopsy images. Although the improvement seems small, the contribution of CNNs was in the important classes of necrosis, epithelium and stroma especially for distinguishing and classifying DCIS. The differentiation between necrosis and secretion might be especially critical in diagnosis. Furthermore, none of the quantitative measurements evaluated the smooth object

boundaries obtained by the CNNs. Because the CNN-based methods were trained with patches that are 500 $\times$ bigger than a superpixel, they were able to learn the structure and segment smooth borders of the objects. This characteristic of the results can be seen in visualizations in Figure 5.12. In the next Chapter, we attempt to develop object-based features for diagnostic classification; having separate objects segmented could be important in the extraction of these features.

Effect of Training Set

In Section 5.3, we trained superixel-based models using a sampling of superpixels from *all* images. In preliminary experiments not reported here, we found this to be a more effective strategy than dividing the dataset into training and test sets. The most important reason for this is the variability in the appearance and the small size of the labeled dataset (N=58). The subset of 58 ROIs was selected using clinical variables and attempted to represent the full digiPATH dataset (N=428). However, due to the diverse nature of the full digiPATH dataset, the subset might have only one or two images representing a certain appearance. In this case, the models trained by an image-split would not see any examples of some images from the test set during training. This might be the case for our CNN-based models since they were trained and evaluated by training and test sets split on images.

Furthermore, we trained superpixel-based models in a 10-fold-cross-validation setting, which allowed us to evaluate the models on all 58 images. However, due to the long training times required by CNNs, this was not possible for the CNN-based models introduced in Section 5.4. In our evaluations, we used only the test set for both CNN-based and superpixel-based models. Although we split the images into training and test sets using clinical variables, the test set was ultimately smaller than that of Section 5.3.

Secretion and Necrosis Labels

Secretion and necrosis labels (and also blood when the 8-label scheme was used) were the rarest labels in the dataset. In the superpixel-based method of Section 5.3, we sub-sampled the other tissue labels during the training of the SVM. Although this strategy ensured the classifier learned

to segment these rare classes, it also caused a lot of false positives for these classes. CNN-based methods, on the other hand, used an augmented dataset that preserved the initial distribution of the labels. This fact may have impacted the performance of the CNN-based methods for these rare classes, because they had the information about their rarity. It may be another reason for the high F_1 score of the CNN-based methods for secretion and necrosis labels in comparison to superpixel-based method.

Labeling Errors

The ground truth labels provided by a pathologist were not perfect. Pixel-wise labeling is an exhaustive job and our dataset was quite big. It is expected from any labeler to change strategies when tired, like using bigger brushes to paint large areas, ignoring small regions inside a big one, or mislabel some border-line regions. Since our evaluations were pixel-based, we acknowledge that no computational model can achieve a perfect segmentation.

5.6 Summary

Tissue-label-based segmentation of breast biopsy images is a useful abstraction of the structures of interest. Structural changes play an important role in the diagnosis of pre-invasive lesions of the breast, including benign proliferations, atypia and DCIS. We developed a set of seven tissue labels to represent the ROIs from the digiPATH dataset. A pathologist labeled all the pixels in 58 ROIs obtained from a subset of 40 WSIs selected from the full dataset preserving the diagnostic category distributions.

We developed a superixel-based methodology to segment the images. Based on our previous work [72], we used color and texture histograms to train an SVM to classify superpixels into eight tissue label classes. We showed that using circular neighborhoods around the superpixel, it is possible to improve the segmentation. The model with two circular neighborhoods achieved an average F_1 score of .35 over 58 ROIs in a 10-fold-cross-validation setting.

Inspired by the effect of context and the recent achievements made by convolutional neural nets, we also developed CNN-based models for segmentation. We used a five-label scheme by

merging two epithelium and two stoma classes, and adding blood label to the stroma class. We compared the superpixel-based model, a Plain-Network, a Multi-Resolution Network and a state-of-the-art Multi-Resolution Multi-Path Network with input-aware encoding blocks. The four methods achieved .28, .23, .34 and .39 F_1 scores, respectively, on the test set of 28 ROIs. We showed that by aggregating features from multiple resolutions and including bigger context, CNNs were able to learn the structures and produce better segmentations.

The segmentation of the breast biopsy slides is not useful in itself but it is an important first step for the automated diagnosis pipeline we introduce in Chapter 6. Based on the segmentation mask, we can describe the structural changes and design image features for automated diagnosis.

Chapter 6

AUTOMATED DIAGNOSIS OF REGIONS OF INTEREST

6.1 Introduction

Diagnostic errors are alarmingly high for pre-invasive lesions of the breast. A recent study shows that the agreement between pathologists and experts for the atypia cases is only 48% [34]. In addition to the benefits of an automated system that can help pathologists during interpretation of slides, studying the visual characteristics of pre-invasive lesions may help identify more discriminative features for diagnosis.

6.1.1 Related Literature

Automated malignancy detection is a well-studied area in the histopathological image analysis literature. Most of the related work focuses on the detection of *cancer* in a binary classification setting with only malignant and benign cases [20, 30, 106]. These methods do not take pre-invasive lesions or other diagnostic categories into account, which limits their use in real-world scenarios. There is also limited research on analyzing images for subtype classification [57] or stromal development [99] using only tumor images.

Recently, some researchers have begun to study the pre-invasive lesions of the breast: [28] reports promising results in discrimination of benign proliferations of the breast from malignant ones. They extract 392 features corresponding to the mean and standard deviation in nuclear size and shape, intensity and texture across 8 color channel, and apply L1-regularized logistic regression to build discriminative models. Their dataset contains only usual ductal hyperplasia (UDH), which maps to benign in the digiPATH data, and ductal carcinoma *in-situ* (DCIS) cases. To our best knowledge, there is no study that considers the full spectrum of pre-invasive lesions of the breast including atypia (atypical ductal hyperplasi and atypical lobular hyperplasia).

6.2 Image Features for Diagnostic Classification

To describe each ROI as a feature vector, we used three levels of features with different configurations:

- *Low-Level Features* : Color and Texture Histograms
- *Mid-Level Features* : Superpixel Label Frequency and Co-occurrence Histograms
  - k-means-Segmentation-Based
  - SVM-Segmentation-Based
  - CNN-Segmentation-Based
- *High-Level Features* : Structure Feature Histograms
  - SVM-Segmentation-Based
  - CNN-Segmentation-Based

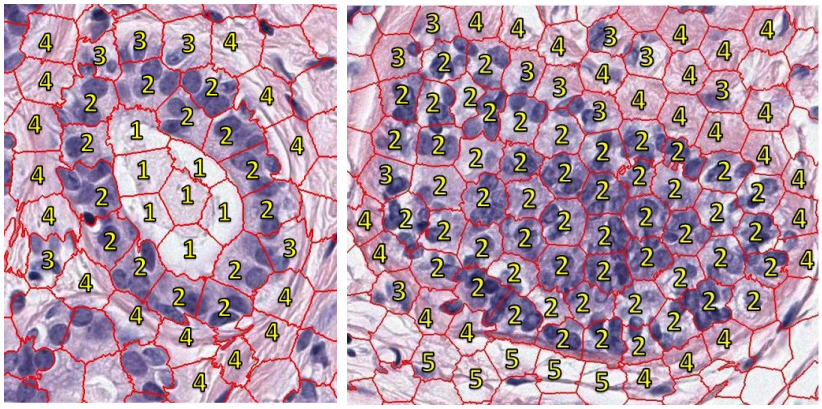
6.2.1 Low-Level Image Features

Low-level features were calculated at every pixel in the ROI and summarized as histograms. In our previous work, we found color and texture to be important features in breast biopsy images [71, 72]. We calculated L*a*b* and H&E color histograms, and LBP texture histogram of H&E channels. The histograms were of size 64 for each channel and were concatenated to produce a feature vector of size $64\times(3+2+2) = 448$.

Low-level features were used as a baseline, since they only provide a summary of the color and texture distribution of the whole ROI. Diagnosis of breast cancer and pre-invasive lesions, on the other hand, is based on structural and cellular properties, which require higher-level features than pixel intensities.

6.2.2 Superpixel Label Frequency and Co-occurrence Histograms

One of the basic visual differences between diagnostic categories is the existence (and amount) of different biological structures. To reduce the complexity of the images, we started with a superpixel



(a) (b)

Figure 6.1: A conceptual picture of cluster co-occurrence features. Superpixels in both images are assigned to hypothetical clusters of (1) empty space, (2) epithelial cells, (3) epithelial cells with some stroma, (4) stroma and (5) loose stroma. **(a)** A breast duct with lumen still empty. The *co-occurrence* of clusters 1-and-2 is high. **(b)** A duct with lumen filled with epithelial cells. The *co-occurrence* of clusters 2-and-2 is high.

segmentation and label each superpixel with one of the three methods discussed below: (1) k-means clustering, (2) SVM-based semantic segmentation and (3) CNN-based semantic segmentation. In all three methods, the result is a labeled image, but in method (1) the label is a cluster identifier while in methods (2) and (3), it is a histopathological tissue class.

Assuming a superpixel can be labeled to capture different common structures, the distributions of superpixel labels in different categories would be different. We calculated *frequency histograms* for the superpixel labels. However, only the existence of some superpixels may not be enough to describe complex spatial relationships. *Co-occurrence* of pairs of superpixel clusters, on the other hand, can capture the frequency of contact between superpixels. For example, an "epithelial nuclei superpixel" next to another "epithelial nuclei superpixel" has a different meaning than an "epithelial nuclei superpixel" next to a "stroma superpixel". Figure 6.1 gives a visual example of a co-occurrence distribution. Co-occurrence frequency is calculated for all pairs of superpixel clusters in the image.

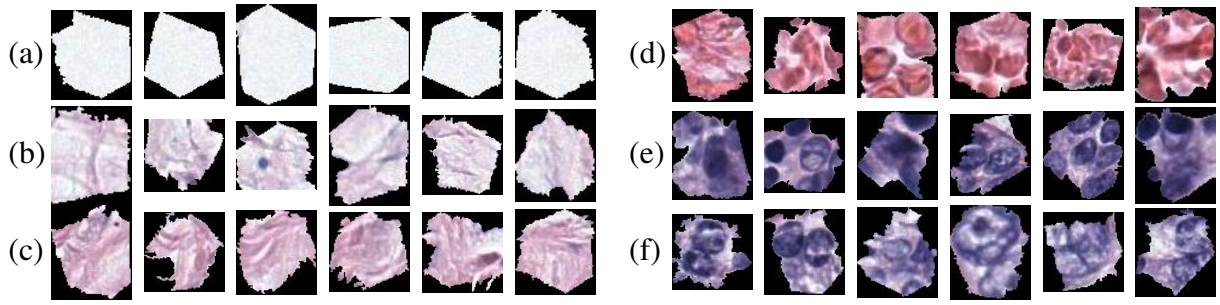


Figure 6.2: Superpixel clusters: Most superpixel clusters can be named by expert pathologists although they are discovered through unsupervised methods. Some of the superpixel clusters as identified by pathologists: **(a)** empty space, **(b)** loose stroma, **(c)** stroma **(d)** blood cells, **(e)** epithelial nuclei, **(f)** abnormal epithelial nuclei.

Unsupervised Discovery of Tissue Types

Similar to our ROI detection method (Chapter 4), we calculated color histograms in L*a*b* color space and texture histograms using LBP [115] for each superpixel. Using k-means clustering, we clustered the superpixels to discover the building blocks of the tissue in an unsupervised way. We found out that most of the superpixel clusters can be identified by pathologists as important structures in the slide. Figure 6.2 shows some example clusters discovered in digiPATH data.

K-means clustering groups superpixels into clusters based on only the appearance (color and texture). When asked to assign each cluster a label, expert pathologists labeled several clusters with the same label. This may be due to the variance in the appearance of certain tissue types; for example, stroma may have distinctly different patterns because of the amount or orientation of the connective tissue fibers. These changes picked up by the clustering algorithm are not biologically significant. Pathologists do not differentiate between these different looking stroma, since they have prior knowledge about the variability of different components in the tissue.

Furthermore, k-means clustering assumes all features have the same spherical variance and the prior probability for all clusters are the same, i.e. each cluster has roughly an equal number of samples. In our case, these assumptions do not hold. Breast biopsy images do not have a uniform distribution of the tissue types we want to segment. Certain tissue labels, like necrosis or blood,

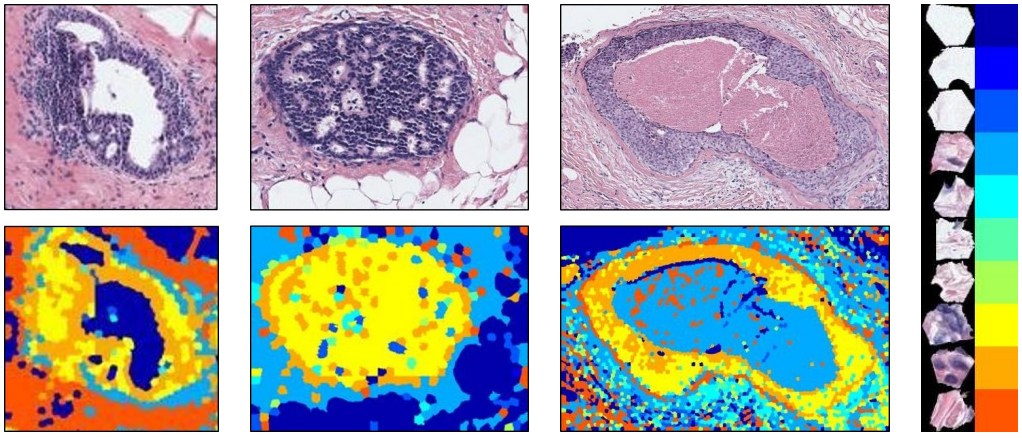


Figure 6.3: **Superpixel cluster visualizations -** The superpixels are colored according to the cluster to which are assigned in three different images from the digiPATH dataset. Top row shows original RGB images and the bottom row shows the visualizations. In this image, (a) is diagnosed as benign, (b) is diagnosed as *atypia* while (c) is diagnosed as *DCIS*. (d) shows 10 clusters and the color codes assigned to them.

are rare, while others like stroma and epithelium are abundant. Thus, k-means divides the bigger clusters of stroma and epithelium into several smaller clusters; resulting in similar looking clusters. Another consequence of non-uniform distribution of tissue types is that rare tissue types cannot be detected when a small k is used for the number of clusters.

Unsupervised segmentation using k-means clustering produced promising segmentations; however, without ground truth segmentations, it was not possible to evaluate the results quantitatively. The visualizations using a color scheme based on the distance between the clusters demonstrates the strength of color and texture features together with superpixels. The qualitative evaluation of the visualizations by the experts motivated the supervised setting described in Chapter 5, which required expert to label a set of the images with tissue labels.

Despite the limitations of unsupervised segmentation, the histogram of tissue clusters can still be powerful as an image descriptor. In contrast to supervised segmentations, unsupervised methods have the potential of discovering new patterns in the images. The subtle color and texture differences between clusters may be deemed insignificant by the pathologists but they may play a role in identifying high-level patterns in the images. Figure 6.3 shows example segmentations.

Supervised Segmentation of Tissue Types

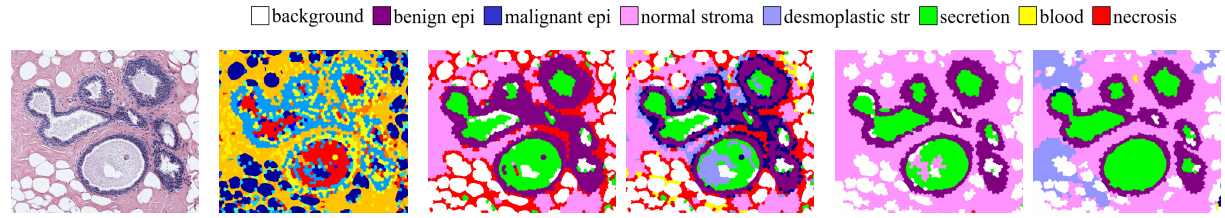
In Chapter 5, we introduced a segmentation framework for the breast biopsy images. A set of seven tissue labels was described: benign epithelium, malignant epithelium, normal stroma, desmoplastic stroma, secretion, blood and necrosis. A total of 58 ROIs from all four diagnostic categories was labeled by a pathologist. Two models were developed to segment the ROIs images into tissue labels. We used the best performing models to segment the full digiPATH dataset of 428 ROIs.

As explained in Section 5.3.3, cross-validation was used to evaluate the performance of different feature combinations using SVMs. Each fold resulted in an SVM model trained on nine tenths of the dataset. For this reason, we trained new SVMs using the superpixels with the two circular neighborhoods (Section 5.3) using all the superpixels from 58 ROIs that were marked by a pathologist. We used the same subsampling technique described in Section 5.3.3 to produce a uniform distribution of tissue labels. We trained two SVMs, one with five tissue labels (background, epithelium, stroma, secretion and necrosis) and one with eight tissue labels (background, benign epithelium, malignant epithelium, normal stroma, desmoplastic stroma, secretion, blood and necrosis). The resulting two SVM models were applied on all ROIs from the full digiPATH dataset (N=428).

Since training CNNs take a lot longer than training SVMs, only one CNN model (MR-MP-Network) was trained with the 30 ROIs for each configuration (Section 5.4). We used two MR-MP-Networks, one trained on five classes and one trained on eight classes; and segmented all ROIs from the full digiPATH dataset (N=428). Because the CNN-based methods classified pixels, not superpixels; we assigned each superpixel the label of the majority of its pixels.

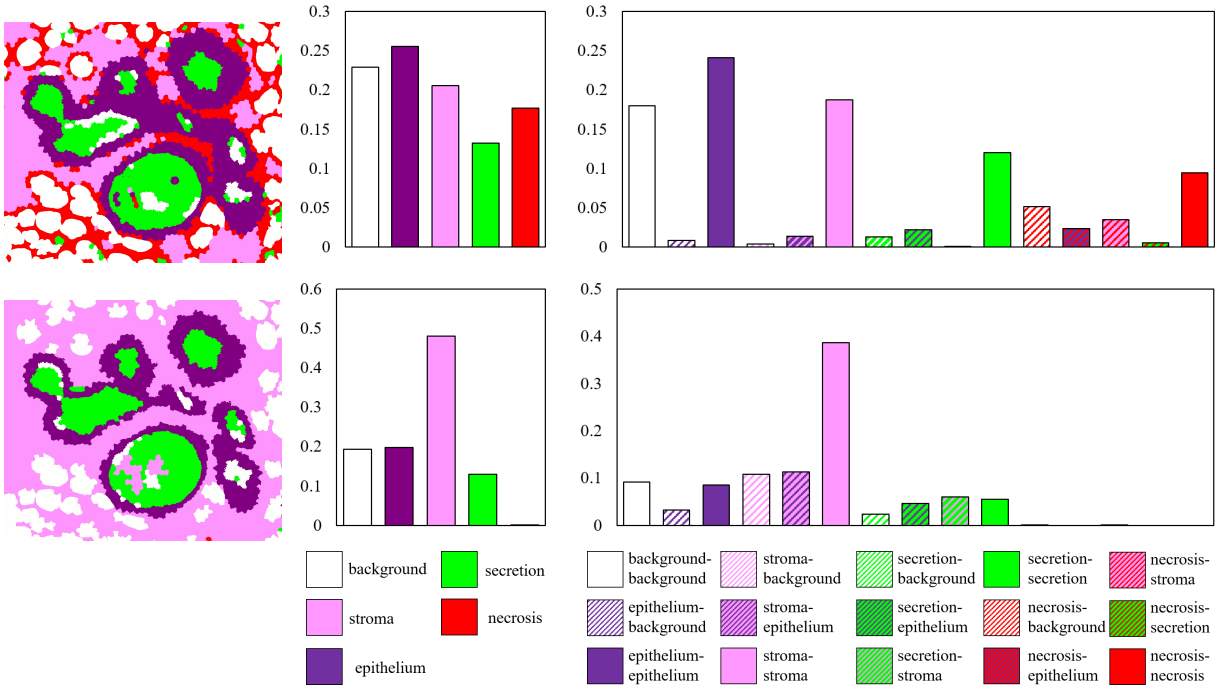
At the end, we obtained four semantic segmentations for all ROIs (N=428):

- *SVM5:* Five-label segmentation using SVM (Figure 6.4c),
- *SVM8:* Eight-label segmentation using SVM (Figure 6.4d),
- *CNN5:* Five-label segmentation using CNN (Figure 6.4e),
- *CNN8:* Eight-label segmentation using CNN (Figure 6.4f).



(a) RGB Image (b) k-means (c) SVM5 (d) SVM8 (e) CNN5 (f) CNN8

Figure 6.4: Unsupervised and supervised segmentation for tissue types: (a) RGB Image, (b) segmentation with k-means clustering (each color is a different cluster but color assignments are arbitrary), (c) supervised segmentation into five tissue labels using an SVM, (d) supervised segmentation into eight tissue labels using an SVM, (e) supervised segmentation into five tissue labels using a CNN, (f) supervised segmentation into eight tissue labels using a CNN.



(a) Segmentation (b) Superpixel Label Frequency Histogram (c) Superpixel Label Co-occurrence Histogram

Figure 6.5: Frequency (b) and co-occurrence (c) histograms for two segmentations (a) of the same image. The errors in the segmentation is accumulated in the histograms.

6.2.3 Structure Feature

We developed an image feature to describe the structural changes that occur at the ducts. A normal breast duct has an empty cylinder shape made up of two layers of epithelial cells. Depending on the slide preparation and the cross-section the image is produced from, the most common appearance of the ducts in the breast biopsy images is an empty circle with two layers of cells. The changes in the duct structure is caused by the proliferation of the epithelial cells. The degree of structural change is an important factor in diagnosis.

Our structure feature describes the changes in the shape of the epithelial structures in the breast biopsy slides. The structure feature is calculated over the tissue label image produced by any supervised segmentation algorithm (SVM5, SVM8, CNN5, CNN8). Using the epithelium labels, we identify objects of interest, which may be a duct, a group of ducts or a tumor. Then starting from the outer border of the object, we extract several layers towards the inside the object and several layers towards the outside. For each layer, we calculate a frequency histogram of the tissue types. Figures 6.6 and 6.7 show two examples of the structure feature extraction on different images.

In our implementation, we used superpixels as the structural elements. The layers are defined 1-superpixel thick and the layer-histograms are built by counting superpixels. Although other structural elements can be used (patches, pixels, hexagons etc.), superpixels provide a good definition for the object borders and are widely used in segmentation. Furthermore, they reduce the size by 3000 pixels to 1 superpixel.

The definition of a layer and the number of layers can be adapted for different datasets and different problems. In our implementation, we defined the layers starting from the outer border of the objects of interest: ducts. We used five inner and five outer layers. Since the size of the superpixels was selected based on the size of an average epithelial cell in our images, the first one or two layers of epithelial superpixels at the circumference of the duct would define a normal duct. Since our dataset had examples from four diagnostic categories, five inner layers and five outer layers were generalizable to all diagnostic categories, yet still powerful enough to describe the structural changes.

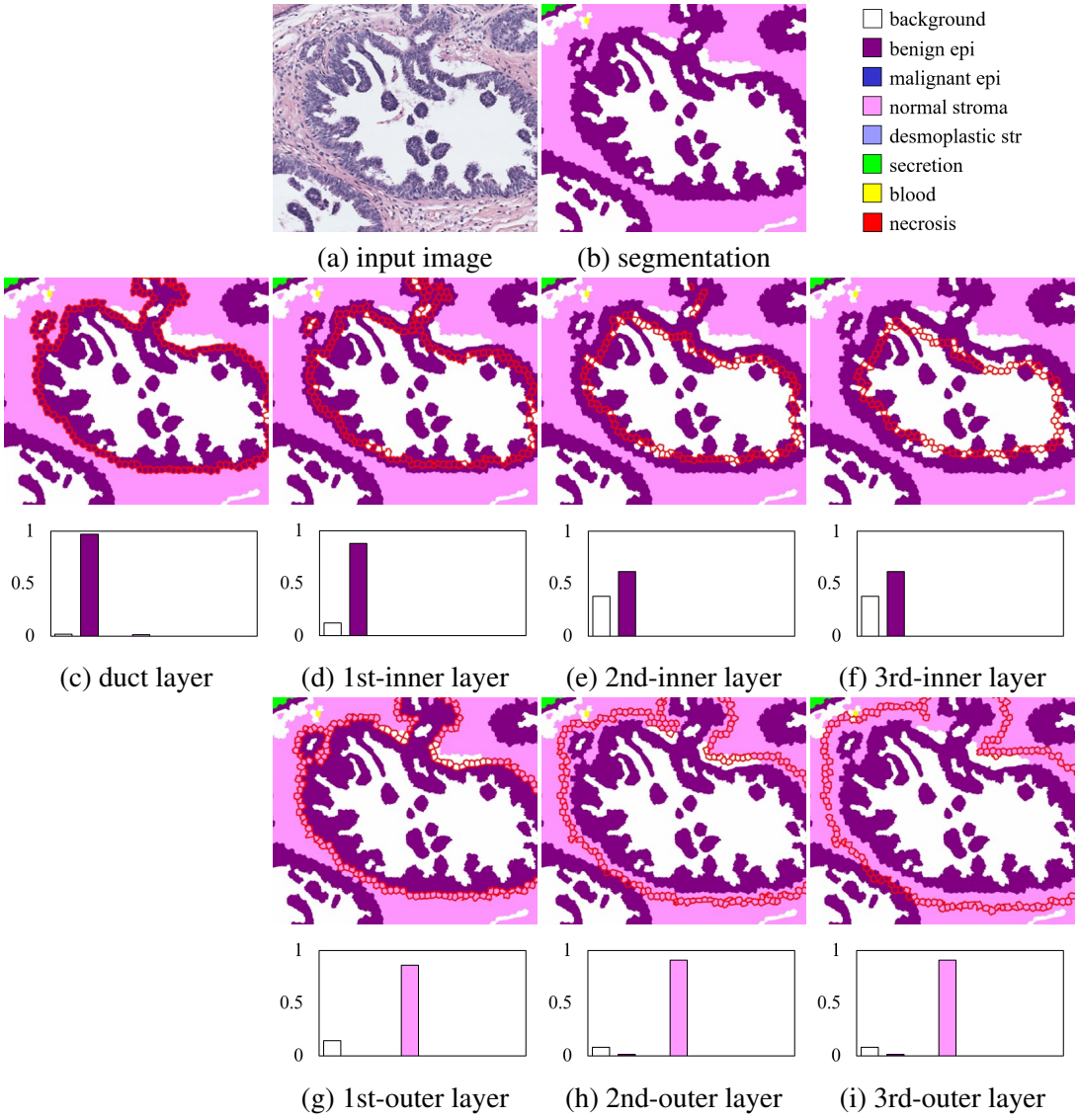


Figure 6.6: An example extraction of the structure feature: Starting with the tissue label segmentation (b), we use epithelium labels as the object of interest. The superpixels at the border of the duct)are used to construct the first histogram for duct layer (c). (d-f) show the first three inner layers and tissue histograms extracted from them while (g-i) show the first three outer layers. The superpixels belonging to a layer are marked with red borders in (c-i).

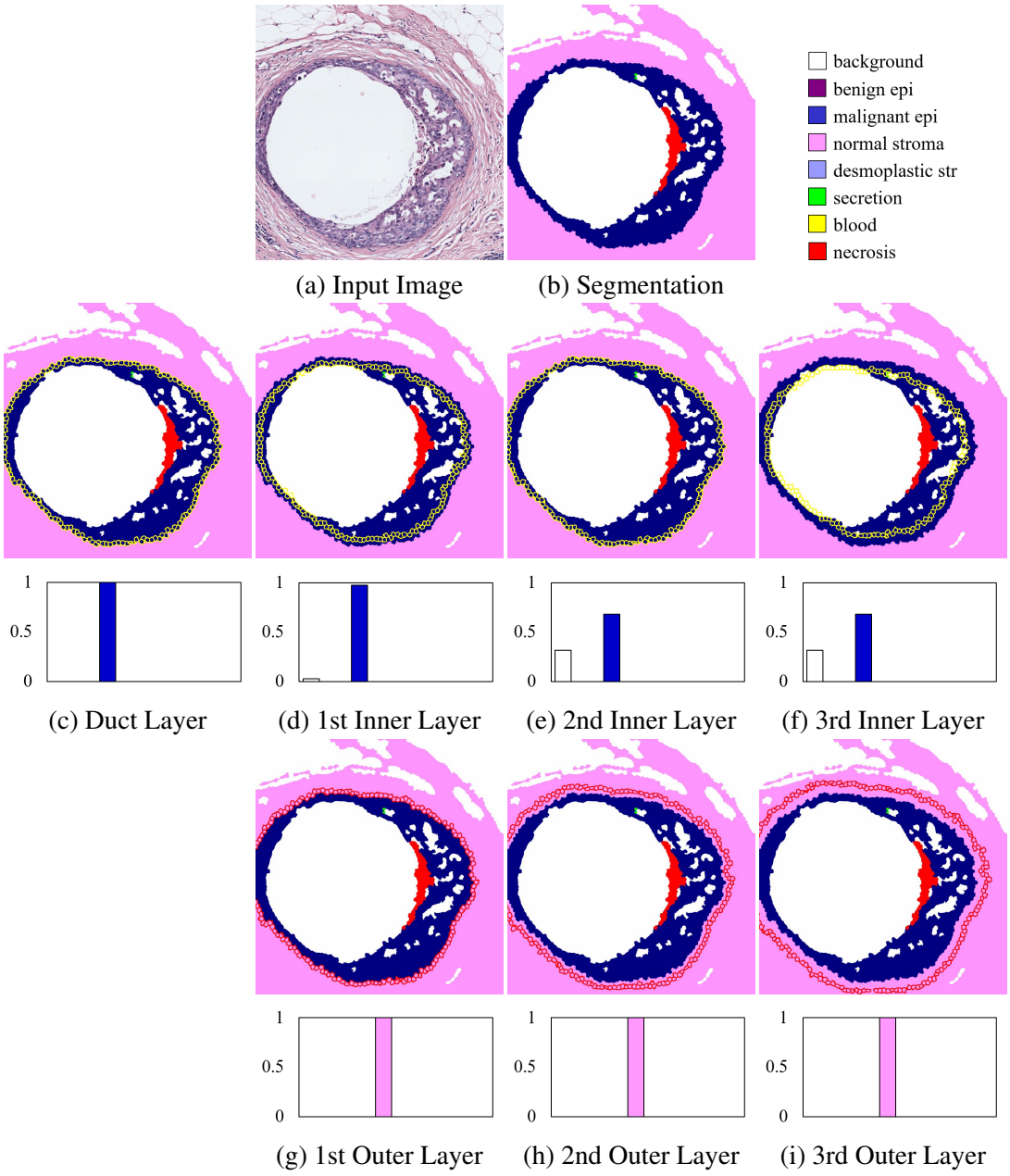


Figure 6.7: An example extraction of the structure feature: Starting with the tissue label segmentation (b), we use epithelium labels as the object of interest. The superpixels at the border of the duct are used to construct the first histogram for duct layer (c). (d-f) show the first three inner layers and tissue histograms extracted from them while (g-i) show the first three outer layers. The superpixels belonging to the duct layer and the inner layers are marked with yellow borders in (c-f), the superpixels belonging to the outer layers are marked with red borders (g-i).

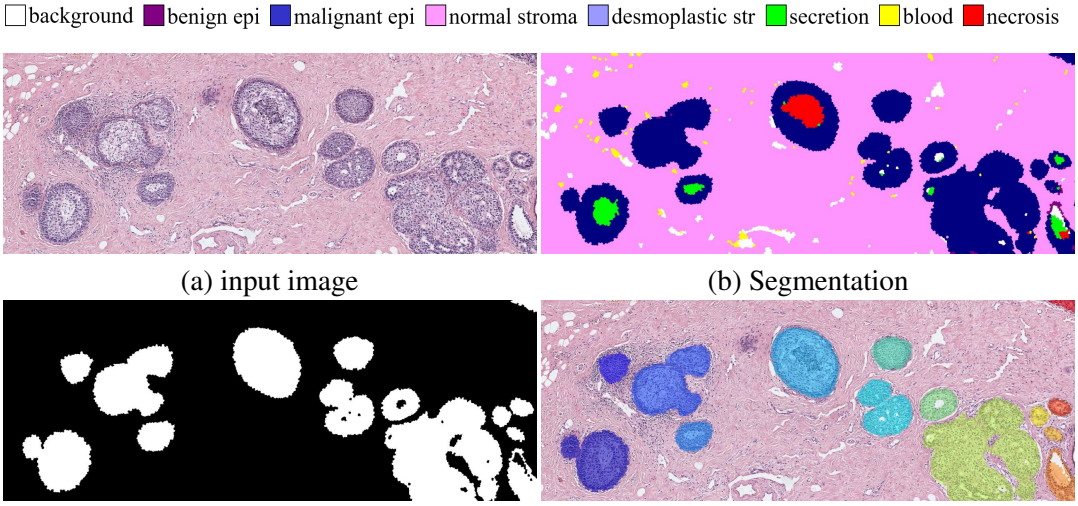


Figure 6.8: Starting with the tissue label segmentation (b) of the image (a), we used epithelium, secretion and necrosis labels as the initial mask (c) and apply connected component analysis (e) to identify the objects of interest.

We used four different tissue label segmentations described in Section 6.2.2: SVM5, CNN5, SVM8 and CNN8. For the segmentation SVM5 and CNN5, the tissue label histograms of size five were calculated for each layer. For the segmentations SVM8 and CNN8, the histograms had eight bins. All segmentation methods introduced some error to the diagnostic classification before the extraction of the structure feature. To reduce some of this error, we excluded any object smaller than three superpixels from the analysis.

Our objects of interest were primarily the breast ducts; however, in certain cases, the duct was filled with secretion or necrosis. To get a complete picture of the structure, we used epithelium (both benign and malignant), secretion and necrosis labels to obtain a binary image. After cleaning small objects, we applied connected components analysis to identify individual objects (ducts or duct groups). Figure 6.8 shows an example image with its tissue label segmentation, the binary image of the union of epithelium, secretion and necrosis labels, and the detected objects overlaid on the original image.

The structure features were calculated for the ducts, but the diagnostic classes were assigned

to the ROIs. In order to obtain a feature vector for each image, we summed up the histograms for each layer. Ideally, we would like to classify each duct but the experts marked the smallest possible region exhibiting the characteristics of the diagnosis assigned to the case. However, some ROIs contained benign ducts or lesser diagnoses than the assigned. By summing up the histograms, we averaged the shape of all the ducts. Yet, assuming the main structures in the ROI are the biggest ones, the feature vector should be dominated by them.

6.3 Diagnostic Classification

We designed a set of experiments to test two diagnostic classification schemes: (1) A model that classifies an ROI into one of the four diagnostic categories (Figure 6.9a), (2) A model that is based on the elimination of one diagnosis at a time (Figure 6.9b).

The diagnostic decision making process is complex. Pathologists interpret the slides at different resolutions and make decisions about different diagnosis. For example, the decision to diagnose an invasive carcinoma is usually made at a lower resolution, where a high-level organization of the tissue is available to the observer. On the other hand, the decision between atypia and DCIS is made at a higher resolution by examining structural and cellular changes (see Chapter 8 and Table 8.3 for more details). Inspired by this observation, we designed a classification scheme where a decision is made for one diagnosis at a time.

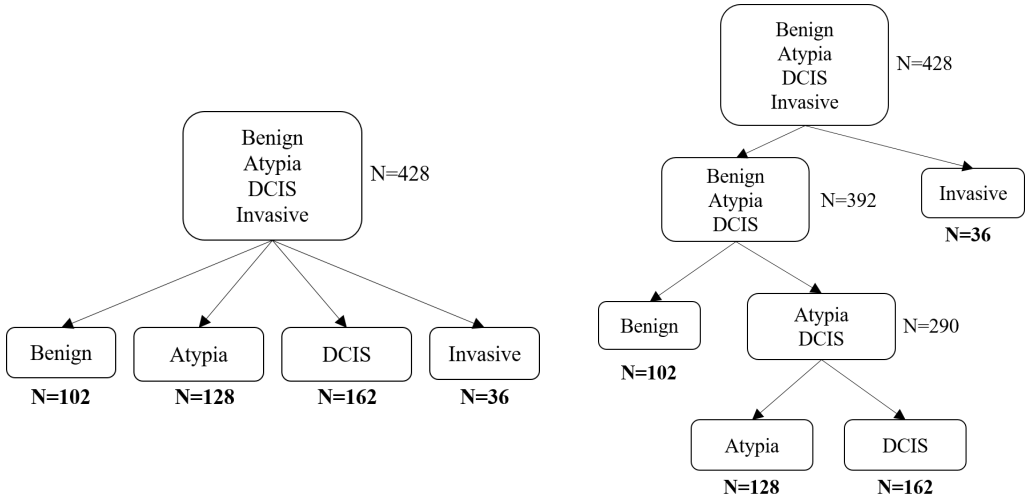


Figure 6.9: Two diagnostic classification schemes: (a) The naive 4-class classification and (b) The smarter classification scheme based on the histopathological decision making process that separates one diagnosis at a time possibly using different features.

6.4 Experiments and Results

6.4.1 Features

We tested all features (low-level, mid-level, high-level) individually for the two classification schemes given in Figure 6.9. For the features based on segmentation, we tested four supervised segmentation models: SVM5, SVM8, CNN5 and CNN8.

Mid-level frequency and co-occurrence feature and high-level structure features consist of histograms. We normalized all histograms to remove the effect of size. Since the background was one of the tissue labels, the amount of background affects the histogram bins of other tissue types, yet the amount of background may not be important in diagnostic classification. We created two alternative versions to all mid-level and high-level features by removing the background bin, and then also removing the stroma bin from the histograms before normalization.

6.4.2 Experimental Setup

For the four-way classification, we trained an SVM using all samples. For the second classification scheme, we trained three SVMs: invasive vs. not-invasive, using all samples; atypia and DCIS vs. benign, using benign, atypia and DCIS samples; DCIS vs. atypia, using atypia and DCIS samples. When the sample size was smaller than the number of features, we applied principle components analysis (PCA) and used the first 20 principal components to reduce the number of features. For all experiments, we trained SVMs in a 10-fold-cross-validation setting. We sub-sampled the training data to have an equal number of samples for each class. To reduce the effect of sub-sampling, we repeated all experiments 100 times and reported the average accuracies.

6.4.3 Dataset

We used consensus ROIs and consensus diagnoses according to the alternative mapping as described in Chapter 3. Our dataset consisted of 102 benign, 128 atypia, 162 DCIS and 36 invasive ROIs.

6.4.4 Results

Table 6.1 shows the classification accuracies for the mid-level and high-level features based on different segmentation methods and different tissue label combinations. We can make following observations from the table:

- None of the methods do well with 4-class classification task. The features with the highest accuracy, 0.56, are the structure features with all tissue labels using the CNN8 segmentation.
- Excluding the background bin from the histogram generally improves the accuracy of both the mid-level and the high-level features for all segmentation models in the differentiating of invasive cancer. However, excluding the stroma bins decreases the classification accuracy for invasive cases.
- In the classification of the atypia and DCIS cases from the benign cases, the structure feature without the background bin using the SVM5 and SVM8 segmentations achieved the best

accuracy of 0.70. Unlike the other classification tasks, the features using CNN8 segmentation did not produce the best accuracies.

- The superpixel label frequency and co-occurrence histograms perform better than the structure feature for the differentiation of invasive cancer. The superpixel label frequency and co-occurrence histograms without the background using CNN8 segmentation achieves the highest accuracy of 0.94.
- For the classification of DCIS vs. atypia, excluding the background bin from the histograms improves the accuracies of CNN-based features. Excluding the stroma bin improves the accuracies further.
- For the differentiation of DCIS from atypia, the structure features perform better than the superpixel label frequency and co-occurrence histograms on average for all segmentation models. The highest accuracy of 0.85, is achieved by the structure feature without background and stroma bins using the CNN8 segmentation.

Table 6.2 compares the *best* accuracies for the structure feature and the supervised superpixel label frequency and co-occurrence features to the low-level and unsupervised superpixel label frequency and co-occurrence features. The superpixel label frequency and co-occurrence features with unsupervised segmentation do not perform as well as their supervised counterparts. Yet, they perform decently for the differentiation of invasive cases by achieving 0.77 accuracy. Low-level color and feature histograms do not perform over random chance but they provide a good baseline showing that more powerful features like the structure features are necessary for diagnostic classification.

6.4.5 Sensitivity and Specificity

Sensitivity and specificity are two metrics that measure the true positive and true negative rates of a condition. They are widely used in the evaluation of diagnostic tests. Sensitivity quantifies the absence of false negatives while specificity quantifies the absence of false positives. Appendix C gives the precision, recall, sensitivity and specificity values for all features in Table 6.1.

Table 6.1: Comparison of mid-level and high-level features with different tissue histograms and segmentations for the diagnostic classification tasks.

Feature	*Segmentation*			
	SVM5	*SVM8*	*CNN5*	*CNN8*
4-class: Benign vs. Atypia vs. DCIS vs. Invasive				
Superpixel label freq. & cooc. hist.	.28	.32	.30	.46
Superpixel label freq. & cooc. hist. (no bg)	.26	.39	.32	.44
Superpixel label freq. & cooc. hist. (no bg or str)	.22	.33	.24	.42
Structure feature	.32	.46	.37	**.56**
Structure feature (no bg)	.32	.41	.37	.55
Structure feature (no bg or str)	.36	.42	.25	.47
Invasive vs. Benign-Atypia-DCIS				
Superpixel label freq. & cooc. hist.	.55	.48	.68	.82
Superpixel label freq. & cooc. hist. (no bg)	.37	.62	.81	**.94**
Superpixel label freq. & cooc. hist. (no bg or str)	.26	.54	.33	.69
Structure feature	.36	.66	.74	.91
Structure feature (no bg)	.36	.62	.76	.91
Structure feature (no bg or str)	.62	.65	.36	.67
Atypia-DCIS vs. Benign				
Superpixel label freq. & cooc. hist.	.61	**.70**	.54	.51
Superpixel label freq. & cooc. hist. (no bg)	.56	.65	.53	.43
Superpixel label freq. & cooc. hist. (no bg or str)	.69	.69	.67	.60
Structure feature	.66	.69	.61	.55
Structure feature (no bg)	**.70**	**.70**	.62	.56
Structure feature (no bg or str)	.66	.69	.66	.61
DCIS vs. Atypia				
Superpixel label freq. & cooc. hist.	.52	.68	.55	.71
Superpixel label freq. & cooc. hist. (no bg)	.57	.65	.57	.78
Superpixel label freq. & cooc. hist. (no bg or str)	.52	.72	.70	.83
Structure feature	.68	.81	.62	.84
Structure feature (no bg)	.62	.78	.56	**.85**
Structure feature (no bg or str)	.65	.75	.69	.83

Table 6.2: Average diagnostic classification accuracies for different features.

	Average Classification Accuracy			
Features	*4-class*	*Invasive*	*Benign*	*DCIS vs. Atypia*
Random Chance	.25	.50	.50	.50
Color and feature histograms	.24	.54	.51	.56
Superpixel label freq. & cooc. hist. with k-means	.36	.77	.65	.60
Superpixel label freq. & cooc. hist. with supervised seg.	.46	**.94**	**.70**	.83
Structure feature	**.56**	.91	**.70**	**.85**

In the classification of the invasive cases, the feature with the highest accuracy, 0.94, the superpixel label frequency and co-occurrence histogram, has a low sensitivity, 0.70, but a high specificity, 0.95. In other words, 30% of the invasive cases were missed, but there were very little false positives. In comparison, the participants had a 0.84 sensitivity and 0.99 specificity.

The atypia and DCIS cases were hardest to classify from benign with only 0.70 accuracy using the structure feature with the SVM5 or SVM8 segmentation and without the background label. The same feature had sensitivity of 0.80 and 0.85, and the specificity of 0.43 and 0.45 for the SVM5 and SVM8 segmentations, respectively. In both cases, almost half the benign cases were overdiagnosed as atypia or DCIS, but the false negative rates were low. In comparison, the participants had 0.72 sensitivity and 0.62 specificity.

The classification of the DCIS vs atypia with the structure feature (without the background) achieved 0.85 accuracy, 0.89 sensitivity and 0.80 specificity for the DCIS cases. The superpixel label frequency and co-occurrence histograms with the CNN8 segmentation, on the other hand, achieved 0.67 sensitivity and 0.93 specificity. There seems to be a trade off between the sensitivity and the specificity for our features; there is not one that achieved the best values in both metrics. In comparison, the participants had 0.70 sensitivity and 0.82 specificity for the atypia cases.

6.4.6 Pathologists' Performance

Table 6.3 shows the confusion matrix for the diagnoses the participants made in the Phase II of the digiPATH using digital slides. To obtain comparable numbers to our classification accuracies, we

Table 6.3: Participants' performance on digital format in Phase II of digiPATH

		Consensus Diagnosis			
		Benign	*Atypia*	*DCIS*	*Invasive*
Participant Diagnosis	*Benign*	933	270	56	22
	Atypia	492	882	375	9
	DCIS	77	182	1386	56
	Invasive	6	1	2	471

took parts of the table and calculated that the accuracy for the invasive cases (vs. benign, atypia and DCIS) was 0.98, the accuracy for the benign cases (vs. atypia and DCIS) was 0.81, and the accuracy for the atypia cases (vs. DCIS) was 0.80.

Our best accuracies for the invasive was 0.91, for the benign was 0.70 and for the atypia was 0.88. Although we did not achieve the participants' accuracies for the invasive and benign cases, we outperformed them for the atypia and DCIS cases.

6.5 Discussion

Classification of regions of interest into diagnostic categories is a crucial evaluation for the image features we are developing. Image features that have high descriptive value are by definition useful in diagnoses and should obtain high accuracy in classification experiments. Once the image features are refined enough, they can be used to describe *any* image region obtained from the analysis of viewport logs.

We designed our structure feature with the changes in the breast ducts of pre-invasive lesions in mind. Because invasive carcinomas have no duct structures left, the superpixel label frequency and co-occurrence features worked better differentiating invasive cases from the others. On the other hand, the structure features were really powerful in classifying atypia cases from DCIS, because the differences in the ducts were captured well by the structure feature.

The classification of benign cases was the most difficult for our features. None of the features got close to the participants' performance. Although the benign cases still had non-proliferative or

proliferative changes differentiating them from the normal breast tissue, the participating pathologists were able to diagnose them.

We used the alternative mapping as our ground truth diagnosis and ROIs (See Section 3.3), which mapped the few ALH and LCIS cases to the atypia and DCIS categories, respectively. Lobular lesions like ALH and LCIS have distinct appearances separating them from ductal lesions. We may not have enough samples of them for the classifier to learn. Furthermore, they may have confused the classifier.

Effect of Tissue Types

We tested structure features and superpixel label frequency and co-occurrence features with and without the background and stroma tissue bins. Removing the background bin helped in general, especially with the CNN8 and CNN5 segmentations. However, removing the stroma bins helped only in benign and atypia vs. DCIS classifications. Stroma, or rather the ratio of stroma, appears to be important in the classification of invasive cases.

Effect of Segmentation

We compared the superpixel label frequency and co-occurrence histograms and the structure features constructed from different segmentations to determine the effect of the segmentation on the diagnostic classification. Not surprisingly, the CNN8 segmentation produced the best results, except for the classification of the benign case. In Chapter 5, we showed that CNNs produced better segmentations than SVMs. The difference may not be much quantitatively but the CNNs produce significantly better segmentations with smooth well-separated ducts.

As expected, the 8-label segmentations (CNN8 and SVM8) outperformed the 5-label segmentations (CNN5 and SVM5). This is due to the diagnostic importance of the malignant and benign epithelium, normal and desmoplastic stroma subcategories. The 5-label segmentation does not have the descriptive power the 8-level segmentation has.

Cellular Features

We only used structural features for diagnostic classification but the pathologists also use cellular features when they make a differentiation between the healthy and cancerous tissues. Incorporating cellular features may improve the classification accuracies, especially for the benign cases.

6.6 Summary

We aimed to develop image features that can describe the diagnostically important visual characteristics of the breast biopsy images. We used low-level features of color and texture histograms and superpixel label frequency and co-occurrence features extracted from k-means clustering and four supervised segmentation models: SVM5, SVM8, CNN5 and CNN8. To describe the structural changes in the breast tissue, we developed the structure feature that describes the tissue label distribution in and around the ducts.

We evaluated out features in classification experiments based on two diagnostic classification schemes: 4-class classification and one-diagnosis-at-a-time classification that starts by separating the invasive cases from the others, then the benign cases from the rest and finally differentiates the atypia cases from DCIS. We found that 4-class classification was too hard to tackle with any of the features, the best accuracy, 0.56. was achieved by the structure features with CNN8 segmentation. For the invasive cases, the superpixel label frequency and co-occurence features with no background and CNN8 segmentation worked the best, achieving a 0.94 accuracy. For the benign cases, the best accuracy was 0.70 which was achieved by the structure feature without the background with both the SVM5 and SVM8 segmentations. We got a 0.85 accuracy for the atypia vs. DCIS classification, using the structure feature without the background and CNN8 segmentation.

Our features could not achieve the participants' performance for the invasive and benign cases, 0.98 and 0.80, respectively; however, they outperformed in the atypia vs DCIS differentiation, 0.80 for the participants. From the experiments, we can conclude that different visual characteristics are important for different diagnostic categories. Our structure feature is especially powerful in describing atypia and DCIS cases. With further improvements in the machine learning tech-

niques used or incorporating cellular level features, it may be possible to get close to the human performance on the automated diagnosis task.

Chapter 7

REGION OF INTEREST IDENTIFICATION AND DIAGNOSTIC CONCORDANCE

7.1 Introduction

The methods and results presented in this chapter were previously published in [76].

7.1.1 Background

Each year, millions of breast biopsies are performed, yet interpreting such specimens is considered to be one of the more challenging areas in pathology. While evaluating a breast biopsy slide, it is critical that the pathologist identifies and then analyzes regions of potential diagnostic interest that might support criteria for diagnosing breast cancer or diagnosing risk-associated non-invasive breast lesions. Pathologists use a complex set of skills to establish a histopathological diagnosis when interpreting a biopsy slide. At least two of these skills may be amenable to computer-assisted image identification and analysis, including: (1) finding relevant diagnostic regions (eg, salient visual features) and (2) interpreting contextual architectural and cytological features in epithelial proliferations.

The fields of pathology and radiology are similar in requiring interpretation of an image to arrive at a diagnosis. Research in interpreting radiology images suggests that searching and diagnosis are possibly separate skills [11]. By inference, insight may be gained by studying how pathologists search for and identify regions of interest (ROI) and then how they diagnose these regions. Digital whole slide imaging in pathology may facilitate this research and lead to future educational and clinical support tools.

In this study, we explored the relationship between areas that pathologists indicated as diagnostic ROIs on whole slide digital images and their diagnostic accuracy. We hypothesized that as the

digitally-marked ROI exhibited increasing overlap with the expert consensus ROI, agreement with the consensus reference diagnosis would increase. This is an intuitive hypothesis for highly reproducible diagnoses such as invasive carcinoma; however, for more ambiguous diagnoses, a 'correct' diagnosis could be based on irrelevant features. Additionally, the method to test this hypothesis is unique and demonstrates the advantages of digital whole slide imaging. Thus, an evaluation of pathologist indicated diagnostic ROI would be informative and novel.

7.2 Quantifying Region of Interest Identification

The participants' diagnoses and marked ROIs were compared with the consensus reference diagnoses and ROIs (Figure 7.1). Diagnostic agreement was defined such that participant diagnoses that agreed with the consensus were given a score of one and those that disagreed, a score of zero. The percent ROI overlap was calculated as the pixel area of the ROI selected by the participant that was within the consensus ROI:

$$\%ROIOverlap = \frac{O}{P} \times 100 \tag{7.1}$$

where P is the number of pixels in the participant ROI selection and O is the number of pixels in the region of overlap (union) between the participant and consensus marked ROI selections (Figure 7.2). Because the consensus ROI could vary in shape, maximum size and number of areas selected, and the participant ROI had a defined maximum size, we defined percent ROI overlap as the proportion of the participant area selected that overlapped with the consensus ROI. This avoids any penalty associated with the participant ROI size restriction.

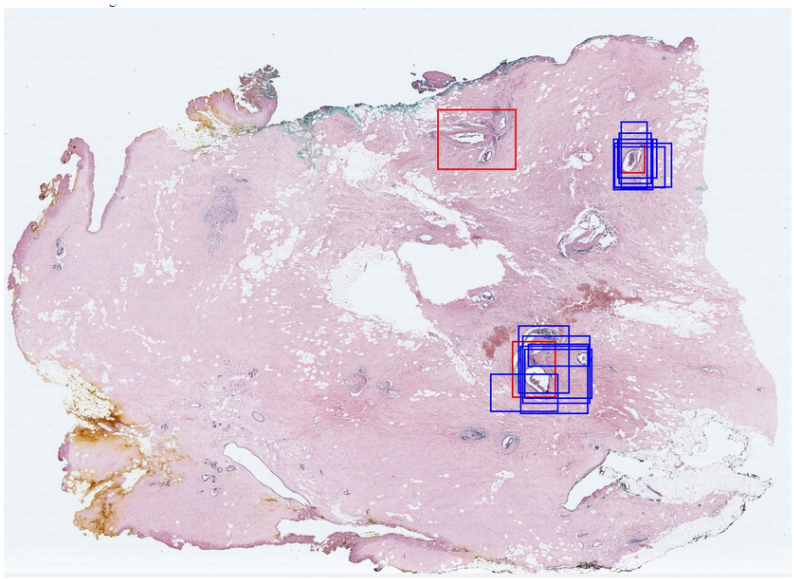
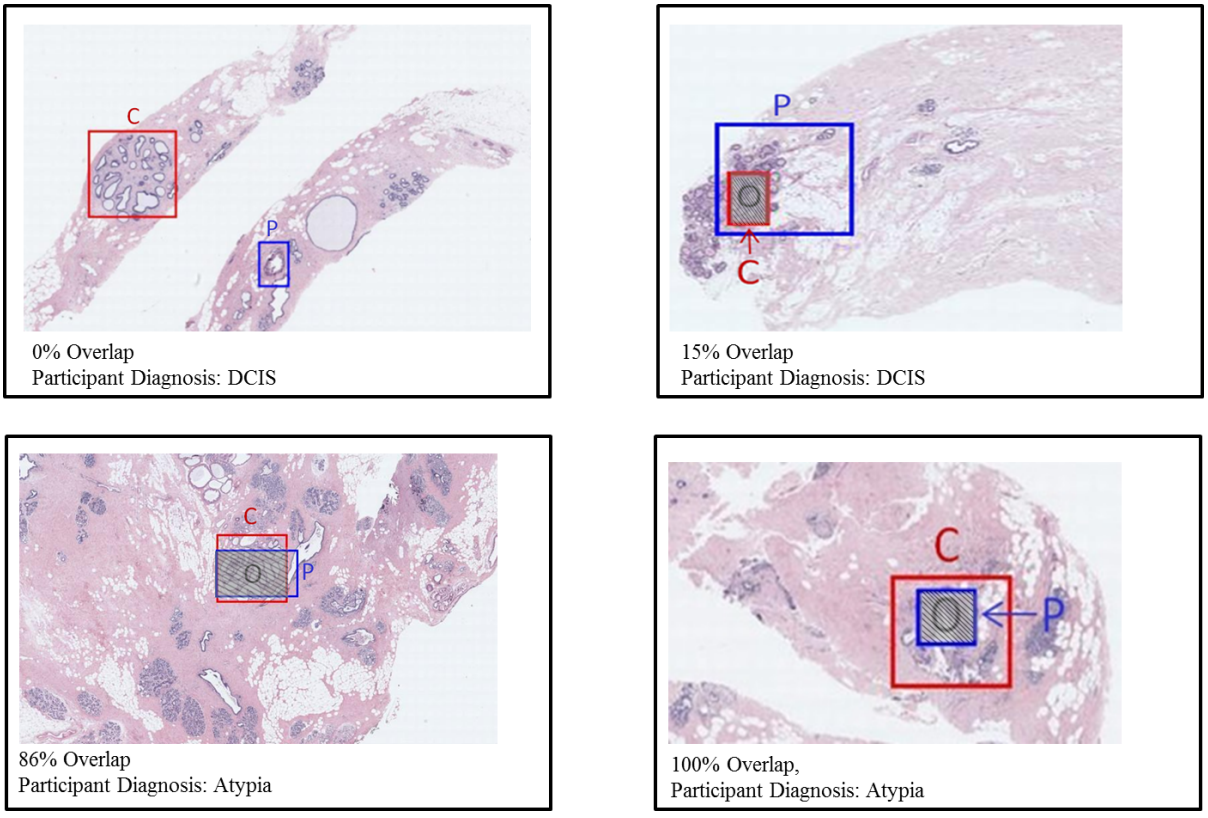


Figure 7.1: Example breast biopsy slide, hematoxylin and eosin stain, demonstrating three different markings for the consensus reference regions of interest (ROI) (shown in red) and the ROI annotations for 12 participants (shown in blue). Participants were instructed to select a single ROI that supported their diagnosis. ©2016, Modern Pathology



%ROI Overlap = (Pixels in O) / (Pixels in P)

O = Overlap
P = Participant ROI
C = Consensus ROI

Figure 7.2: Example cases defined as atypia by the expert consensus with expert consensus ROI markings (in red), participant ROI markings (in blue), and example determination of percent ROI overlap for the case. ©2016, Modern Pathology

7.3 Statistical Analysis

In this study, we used the ROIs, histology form data and baseline survey data from 44 participants as explained in Section 3.4. As the consensus reference ROIs and diagnosis, digital consensus diagnosis based on alternative mapping was used (Sections 3.3 and 3.2). Benign test cases were excluded in this pilot study, as there was often no specific abnormality present on these slides to mark as an ROI, i.e. any ROI on the slide would be 'correct'.

Summary statistics in descriptive tables and regression estimates of percent ROI overlap were obtained using a repeated design model based on generalized estimating equations (GEE) and an independent working correlation structure to account for correlated responses within participants.

Because prominent bi-modality was present at 0 and 100% ROI overlap, and to aid interpretation,the ROI overlap was transformed to a categorized ordinal variable (0, 1–33, 34–65, 66–99, and 100%). We identified covariates for inclusion into the multi-variable analyses based on a criterion of $P < 0.05$. We fitted the outcome and percent ROI overlap model with diagnostic classification as a moderating variable and any covariate that remained significant in the adjusted model at $P < 0.15$. The categorized ordinal variable reported the change in log odds of agreement for each increment of change in ROI overlap.

To produce odds ratios from significant interactions, the ordinal main effect of ROI overlap was re-parameterized by creating a separate ROI overlap variable for each category within each diagnostic class. Odds ratios and their corresponding Wald p-values and 95% confidence intervals were calculated to test the significance of the independent effect of ROI on agreement. A p-value of 0.05 (two-tailed) or a 95% confidence interval not including unity was considered as statistically significant. All statistical analyses were conducted using SAS software version 9.4 (SAS Institute, Cary, NC, USA).

7.4 Pathologist Characteristics and ROI Overlap

Characteristics of the 44 participating pathologists are shown in Table 7.1 along with the percentage ROI overlap for each characteristic. The majority of pathologists were over 50 years of age (68%),

male (73%), and work in a facility with fewer than 10 pathologists (66%) with no affiliation with an academic medical center (75%). The only characteristic associated with higher ROI overlap was academic affiliation ($p = 0.002$).

Table 7.1: Characteristics of pathologists (N= 44) and average percent overlap of the region of interest (ROI) marked by participating pathologists and the ROI marked by the expert consensus (N= 1972 total interpretations) ©2016, Modern Pathology

Pathologist characteristics	*# Pathologists (%)*		*Participant % ROI*		*p-value*
Total	44	(100.0)	66	(64-67)	-
Demographics					
Age at survey (years)					
30-39	6	(13.6)	65	(61-69)	0.92
40-49	8	(18.2)	67	(63-72)	
50-59	19	(43.2)	66	(63-68)	
60+	11	(25.0)	66	(62-69)	
Gender					
Male	32	(72.7)	66	(65-68)	0.38
Female	12	(27.3)	65	(61-69)	
Breast pathology expertise					
Facility size					
<10 pathologists	29	(65.9)	66	(65-68)	0.69
≥ 10 pathologists	15	(34.1)	65	(62-69)	
Fellowship training in surgical or breast pathology					
No	19	(43.2)	67	(65-70)	0.067
Yes	25	(56.8)	65	(62-67)	
Affiliation with academic medical center					
No	33	(75.0)	65	(63-66)	0.002
Yes	11	(25.0)	69	(67-72)	
Do your colleagues consider you an expert in breast pathology?					
No	36	(81.8)	66	(64-68)	0.36
Yes	8	(18.2)	64	(61-68)	
Breast pathology experience (years)					
<5	8	(18.2)	67	(63-70)	0.37
$5-9$	5	(11.4)	69	(65-74)	
$10-19$	17	(38.6)	65	(62-68)	
≥ 20	14	(31.8)	65	(63-68)	
No. of breast cases (per week)					
<5	12	(27.3)	65	(62-67)	0.63
$5-9$	19	(43/2)	66	(64-69)	
$10+$	13	(29.5)	66	(63-70)	

7.5 Case Characteristics and ROI Overlap

Characteristics of the 180 test cases are shown in Table 7.2 along with the percentage ROI overlap for each characteristic. Case characteristics associated with higher ROI overlap included biopsies from women with lower breast density ($p < 0.001$) and disease classification ($p < 0.001$). In addition, cases with higher severity diagnoses (invasive and DCIS) had more ROI overlap than the atypia cases. Additionally, the cases with the lowest number of assessment terms applied by the participating pathologists to the case ($p < 0.001$), with assessments that were rated as very easy or easy to interpret ($p = 0.005$), and with no borderline call ($p < 0.001$) had significantly higher overlap of participants ROI markings with the reference ROI markings.

Table 7.2: Breast biopsy case characteristics and average percent region of interest (ROI) overlap of pathologists participating in the breast pathology study and the expert consensus ROI (n = 1972 independent interpretations by participating pathologists) ©2016, Modern Pathology

Patient and case characteristics	*# Interpretations (%)*		*Participant % ROI*		*p-value*
Total interpretations	1972	(100.0)	66	(64-67)	-
Patient characteristics					
Breast density					
Low density	941	(47.7)	69	(67-71)	< 0.001
High density	1031	(52.3)	63	(61-65)	
Case characteristics					
Biopsy type					
Core needle biopsy	1119	(56.7)	65	(64-67)	0.48
Excisional biopsy	853	(43.3)	66	(64-69)	
Expert consensus diagnosis					
Atypia	871	(44.2)	51	(48-54)	< 0.001
DCIS	859	(43.6)	73	(70-75)	
Invasive	242	(12.3)	93	(90-95)	
Cumulative number of unique individual assessment terms given to a case by participants[a]					
<4	393	(19.9)	85	(83-88)	< 0.001
4–7	1161	(58.9)	64	(63-66)	
≥ 8	418	(21.2)	54	(50-58)	
Level of diagnostic difficulty of this case					
Easy	1378	(69.9)	68	(66-69)	0.005
Challenging	594	(30.1)	62	(58-65)	
Confidence in assessment					
High confidence	1606	(81.4)	67	(65-68)	0.062
Low confidence	366	(18.6)	63	(59-66)	
Case considered borderline					
Yes	524	(26.6)	57	(54-61)	< 0.001
No	1448	(73.4)	69	(67-71)	

[a]Includes 14 terms: non-proliferative changes only, fibroadenoma, atypical lobular hyperplasia, lobular carcinoma in situ, intraductal papilloma without atypia, usual ductal hyperplasia, columnar cell hyperplasia/columnar cell change, sclerosing adenosis, radial scar/complex sclerosing lesion, flat epithelial atypia, atypical ductal hyperplasia, intraductal papilloma with atypia, ductal carcinoma in situ (DCIS), and invasive carcinoma.

7.6 ROI Overlap and Diagnostic Concordance

A total of 1972 individual diagnoses on cases with marked ROIs were available for analyses. The box plot representation of data (Figure 7.3) shows that the median percent ROI overlap for assessments with agreement of diagnoses was 89% (interquartile range: 52%-100%). This was significantly higher than the median percent overlap of 53% (interquartile range: 0%-89%) for assessments with disagreement in the diagnosis ($p < 0.001$ by Friedman test).

A statistically significant positive trend in diagnostic agreement was noted in the aggregate data for all cases when percent ROI overlap was categorized into five incremental groups of increasing overlap (Figure 7.4), ($p - trend < 0.001$). When stratified by diagnostic classification (atypia, DCIS, invasive cancer), the predominantly trend exhibited in the aggregate data was less pronounced, particularly for invasive disease (Figure 7.5). For example, the results for invasive breast cancer suggest a binary relationship; when there was no ROI overlap, the diagnostic agreement was lower rather than an upward linear trend. In invasive cancer cases, there was 75% agreement for cases with no overlap and over 91% agreement for cases with any level of ROI overlap. However, the numbers were small and these findings only suggestive. In contrast to invasive cancer, the relationship between percent ROI overlap and diagnostic concordance for atypia cases suggests a linear relationship. The DCIS cases demonstrate an intermediate pattern between invasive and atypia. For DCIS cases, there was 51% agreement with the reference diagnosis when there was no ROI overlap, but agreement increased from 82% to 90%, as the percent ROI overlap increased.

The unadjusted odds of participant diagnostic agreement with the expert consensus diagnosis increased with each incremental increase in ROI overlap category (Figure 7.6). For example, the unadjusted odds of agreement with the consensus diagnosis was 2.2 times greater (odds ratio [OR] 2.2, 95% confidence interval [CI] 1.5-3.0, $p < 0.001$) when the area of the ROI selected by the participant encompassed between 1% and 33% of the consensus ROI compared to the reference of no ROI overlap. When the ROI overlap was 100%, the odds of agreement in the diagnosis were more than seven times greater (OR 7.7, 95% CI [5.2-11.3], $p < 0.001$) (reference 0% overlap in the ROI). The global p-value for this association was $p < 0.001$. The unadjusted ordinal effect of

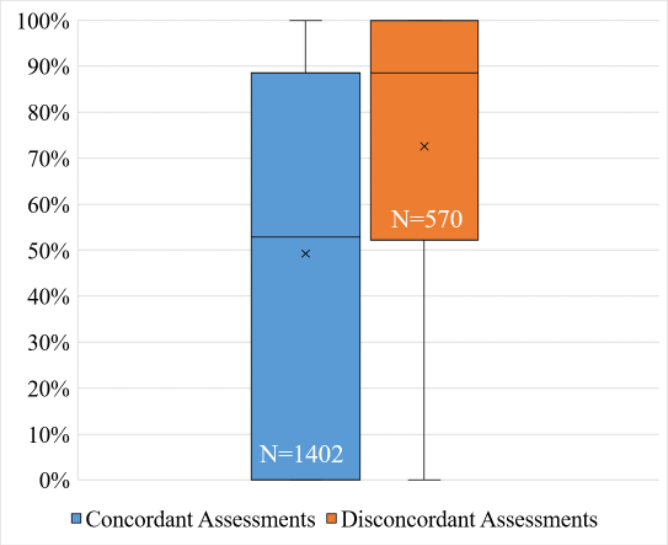


Figure 7.3: Box plot of percent region of interest (ROI) overlap of participating pathologists and the consensus reference ROI by diagnostic concordance or discordance with the consensus reference diagnosis. ©2016, Modern Pathology

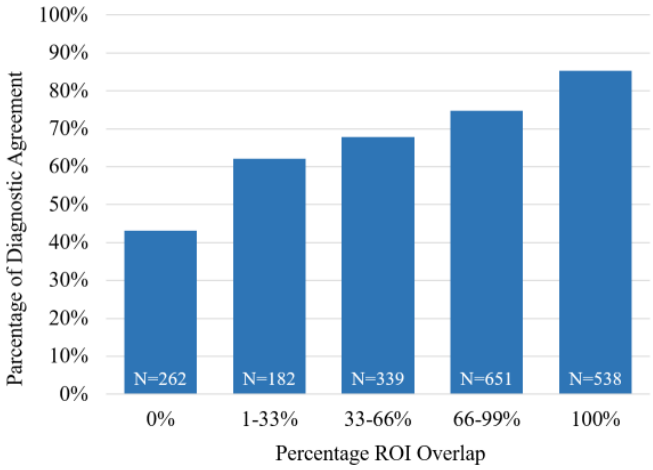


Figure 7.4: Average diagnostic agreement with the consensus reference diagnosis for individual assessments in five categories of percent ROI overlap with the consensus reference ROI (N= 1972 individual assessments). The numeric labels within the bars represent the number of individual assessments within each percent ROI overlap category. ©2016, Modern Pathology

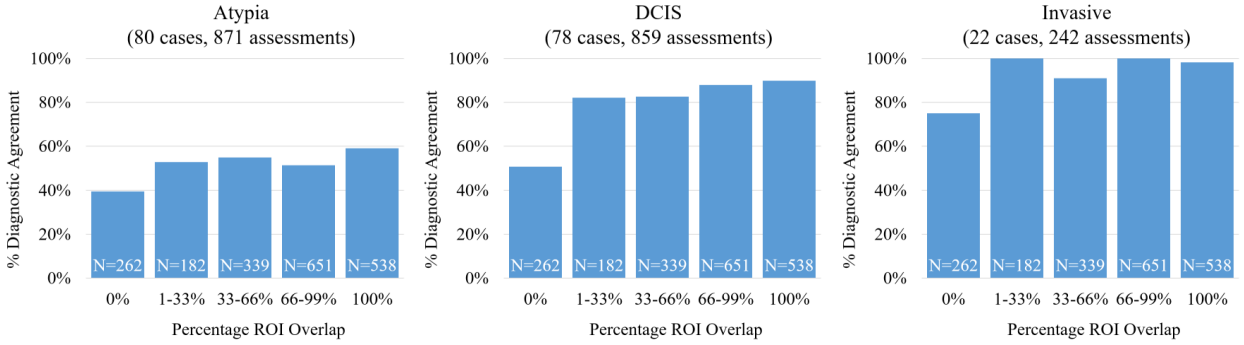


Figure 7.5: Average diagnostic agreement with the consensus reference diagnosis for individual assessments in five categories of percent ROI overlap with the consensus reference ROI (N=1972 individual assessments) shown by reference diagnostic category. ©2016, Modern Pathology

ROI overlap on agreement of the diagnoses increased 60% for each incremental increase in ROI overlap category (OR 1.6, 95% CI [1.5 - 1.7], $p < 0.001$) (data not shown).

The main effects of ROI overlap and diagnostic classification remained significant in the interim model, at the $p < 0.15$ significance level, after adjusting for participant academic affiliation and case characteristics (diagnostic classification, breast density, and case difficulty) and thereby created the final model. The interaction term representing ROI overlap and consensus diagnosis was significant ($p < 0.001$), indicating that the association between ROI overlap and agreement with the reference diagnosis was modified by the consensus reference diagnosis category of the case. For each incremental increase in the ordinal predictor of ROI overlap category, agreement with the consensus reference diagnosis increased 130% when the reference diagnosis was invasive disease (OR 2.3, 95% CI [1.7 - 3.1], $p < 0.001$). For atypia and DCIS, the magnitude of the effect of increasing ROI overlap on concordance with the reference diagnosis was less. The incremental effect of increasing ROI overlap on agreement among DCIS cases was 60% (OR 1.6, 95% CI [1.4 - 1.9], $p < 0.001$). When interpreting atypia cases, the effect of increasing ROI overlap resulted in a 20% increase in agreement (OR 1.2, 95% CI [1.0 - 1.3], $p = 0.01$) (data not shown).

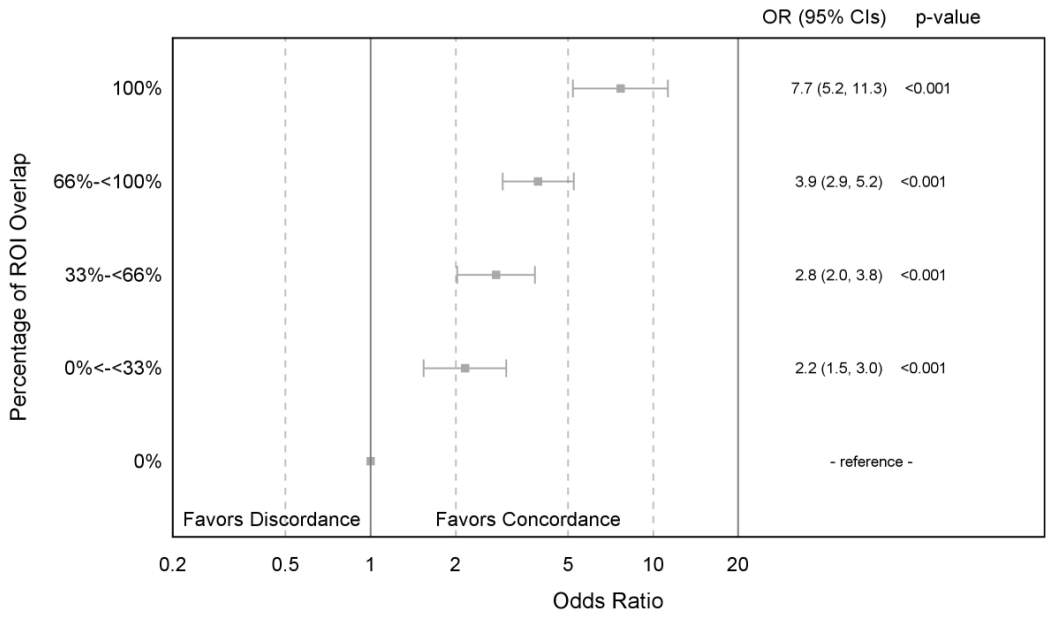


Figure 7.6: Unadjusted odds of diagnostic concordance by increasing ROI overlap category. ©2016, Modern Pathology

7.7 Discussion

Our study investigated whether diagnostic agreement increases as pathologists identify and mark the same region on a slide as annotated by an expert consensus reference standard. In the simplest terms, if two pathologist observers are looking at the same location and features on a slide image, they are more likely to arrive at the same diagnosis. The largest incremental increases in diagnostic concordance were observed between the 0% ROI overlap and the next lowest category (1–33% ROI overlap), especially for the invasive breast cancer cases; smaller incremental increases in diagnostic concordance were observed as percent ROI overlap increased incrementally.

Our findings support the concept that diagnostic accuracy in pathology is dependent, first, on a visual scanning and search process and locating potentially important diagnostic ROIs within medical images, and second, on discriminately focusing on particular diagnostic features within the ROI. These concepts are intuitive to teachers and trainees but are challenging to measure objectively. Our observation that the largest incremental increase in diagnostic concordance occurred

between no observed overlap of the ROI annotations and a small amount of overlap supports the concept that searching for and identifying potential diagnostic features is a critical skill.

Eye-tracking studies of trainee and practicing pathologists have further shown this scanning and targeted focusing behavior becomes increasingly pronounced as trainees gain advanced experience [60, 61]. Digital WSI may provide opportunities to improve the way we train pathologists and how we evaluate histopathology skills, leading to improvements in overall diagnostic abilities and medical care. The inclusion of automated algorithms to highlight ROI may focus a practicing pathologists attention on potentially diagnostic regions, particularly when non-diagnostic distracting features are present. Automated algorithms used in radiology to accentuate relevant radiographic characteristics have proven effective in improving radiology trainee localization of critical diagnostic features [80]. Similarly, masking and unmasking diagnostic ROIs in the field of pathology may help trainees become proficient at identifying diagnostic regions.

The data demonstrate a binary trend for DCIS and invasive cases where any level of ROI marking overlap achieves significantly higher levels of diagnostic agreement with a reference diagnosis. The cases of atypia alone do not demonstrate the same binary pattern. Previous studies have shown that breast atypia is consistently the most challenging diagnostic category for pathologists [6, 34, 93, 98]. Thus, even if a pathologist reviews a critical image region, recognizing architectural and cytologic features of atypia and accurately assimilating these features into a diagnostic rubric can be challenging.

Increased breast density on mammography has been associated with lower sensitivity of the radiologists interpretations [9, 104, 108]. As breast density increases, there may be significantly more background proliferative change for the pathologist to review on the slide, which could make it more difficult to screen for the most relevant ROI. We did note a trend that percent ROI overlap decreased with increasing breast density. However, our test cohort included only 18 cases with the highest density classification, and thus our analysis was restricted to a binary density classification rather than evaluating the four BI-RADS density categories independently. Thus, our data suggest that interpreting the images of denser breast tissue is challenging for both radiologists and pathologists, and further research is indicated.

Technical challenges encountered during this pilot study may help to inform future research and experimental design. Diagnostic agreement of participants was still noted when there was no ROI overlap; values for diagnostic agreement ranged from 40% for atypia cases to 50% for DCIS cases to 75% for invasive carcinoma cases when percent ROI overlap was zero. A non-overlapping participant ROI could theoretically occur for three reasons: (1) the ROI area selected by participants did not meet the reference panel criteria for diagnosis, (2) the reference panel did not include all diagnostic regions within the consensus ROI, and (3) the participant annotated an ROI that did not support their final interpretation. The first two explanations are associated with standard issues associated with diagnostic accuracy: pathologists may have different opinions as to whether a feature set meets diagnostic criteria. The third explanation is a technical limitation. Although the viewer software required annotation of an ROI to complete a case, there was no quality control check to assure that the participants ROI supported their final diagnosis rather than an intermediate conclusion or whether the ROI even contained breast tissue or epithelium rather than a non-relevant part of the slide image.

Another technical challenge was that we allowed the reference ROI to be larger and include multiple shapes compared with the participant ROIs. Thus, a future educational tool might benefit from rank ordering ROIs when multiple reference ROIs are present. The participant ROI had a maximum size limitation and the ROI was constrained to a rectangular shape. Thus, when the reference ROI was smaller than the participant maximum allowed ROI, the percent ROI overlap could be low if the participant did not carefully draw the ROI around only the diagnostic features. This may have limited our ability to conclude that diagnostic accuracy increases as pathologists focus on the same diagnostic features. Since we observed that diagnostic accuracy generally increased as percent ROI overlap increased, we cannot exclude the conclusion that pathologists who were the most discriminating when drawing an ROI tightly bordering the diagnostic features were also the pathologists who most rigorously applied and interpreted diagnostic criteria. Finally, confluent areas of diagnostic invasive carcinoma were generally large, limiting the association between percent ROI overlap and diagnostic concordance for invasive cancer cases. We also did not evaluate the participants full interpretive behaviors using computer image analysis or analyze the image content

on each slide. Future research should include linking such captured diagnostic image features and linking these to patient outcomes.

Other non-technical limitations included the use of only one slide per case, which may not represent clinical practice. However, this could be beneficial in a study scenario, as it limits the amount of variance in interpretation and can help in better isolating specific tissue characteristics that lead to particular diagnoses. Another limitation is that this study was done only in breast tissue, and it is unknown whether these findings would be relevant to other areas in pathology.

Limitations aside, this study is the first of its kind. Strengths of this study include the large number of cases and high number of participants, many of whom spent up to 20 hours participating in the larger study without compensation other than an opportunity to earn continuing medical education credits. Additionally, each test case had a carefully defined expert-based consensus reference diagnosis for comparison and defining accuracy.

In conclusion, this study used digital WSI in a novel manner and demonstrates a potential application of WSI for teaching and improving diagnostic skills of pathologists. Identifying an important region on a histopathology slide image is a significant predictor of diagnostic accuracy and thus may be an indirect indicator of search and screening skills. These findings suggest that computer-aided detection algorithms that highlight potential regions of diagnostic interest on pathology slide images may potentially improve diagnostic accuracy.

7.8 Summary

A pathologist's accurate interpretation relies on identifying relevant histopathological features. Little is known about the precise relationship between feature identification and diagnostic decision making. We hypothesized that greater overlap between a pathologist's selected diagnostic region of interest (ROI) and a consensus derived ROI is associated with higher diagnostic accuracy. We developed breast biopsy test cases that included atypical ductal hyperplasia (n=80); ductal carcinoma in situ (n=78); and invasive breast cancer (n=22). Benign cases were excluded due to the absence of specific abnormalities. Three experienced breast pathologists conducted an independent review of the 180 digital whole slide images, established a reference consensus diagnosis and

marked one or more diagnostic ROIs for each case. Forty-four participating pathologists independently diagnosed and marked ROIs on the images. Participant diagnoses and ROI were compared with consensus reference diagnoses and ROI. Regression models tested whether percent overlap between participant ROI and consensus reference ROI predicted diagnostic accuracy. Each of the 44 participants interpreted 3950 cases for a total of 1972 individual diagnoses. Percent ROI overlap with the expert reference ROI was higher in pathologists who self-reported academic affiliation (69 vs 65%, P=0.002). Percent overlap between participants ROI and consensus reference ROI was then classified into ordinal categories: 0, 133, 3465, 6699 and 100% overlap. For each incremental change in the ordinal percent ROI overlap, diagnostic agreement increased by 60% (OR 1.6, 95% CI (1.51.7), $P < 0.001$) and the association remained significant even after adjustment for other covariates. The magnitude of the association between ROI overlap and diagnostic agreement increased with increasing diagnostic severity. The findings indicate that pathologists are more likely to converge with an expert reference diagnosis when they identify an overlapping diagnostic image region, suggesting that future computer-aided detection systems that highlight potential diagnostic regions could be a helpful tool to improve accuracy and education.

Chapter 8

CHARACTERIZING DIAGNOSTIC SEARCH PATTERNS: SCANNERS AND DRILLERS

8.1 Introduction

The methods and results presented in this chapter were to be published in [73].

8.1.1 Motivation

Pathologic diagnosis is a complex process characterized by visual search and interpretation strategies. Previous research concerning the visual search patterns of physicians has focused on volumetric lung images [8, 16, 31], mammography [109], and breast pathology [26, 61]. The method of investigation has usually included eye-tracking or video recordings of physicians interpreting medical images in a setting controlled by the experimenter. Three outcomes from published research are relevant to the present study. First, physicians reviewing medical images tend to adopt one of two search strategies: drilling versus scanning. Drilling involves restricting a search to a region of interest and zooming in to high magnification levels. Conversely, scanning involves maintaining a particular zoom level while searching relatively broad regions of interest [31]. Second, search strategies change as a function of acquired experience in an expert domain [16, 59] and prior experience with novel review formats [111]. Third, certain visual search strategies have been associated with greater diagnostic accuracy and efficiency. In radiology, physicians who use a drilling search pattern tend to show higher accuracy and efficiency when detecting lung nodules in volumetric images [31, 116], though no research has explored drilling and scanning strategies by pathologists reviewing non-volumetric images.

To address this knowledge gap, our study attempts to provide an initial understanding of the interpretative strategies pathologists use when reviewing digital slides of breast biopsy specimens.

In this study, we investigated three aims. First, we considered how various pathologist characteristics are associated with the two image review strategies (drilling and scanning) identified in the extant cognitive science literature [31]. Second, we tracked how these image review strategies may change as pathologists gain experience with the digital imaging format. Finally, we examined the extent to which each interpretive strategy is associated with diagnostic accuracy and efficiency.

While digital slides are becoming a powerful adjunct tool for breast pathology, understanding the diagnostic processes used by pathologists as they interpret cases may provide insight to improve the education and training of pathologists and lead to the development of computational tools that can aid in the diagnostic decision making process.

8.1.2 Related Literature

Analysis of scanning behavior in histopathology is an emerging research area. Researchers have used eye-tracking systems [16, 60], think-out-loud protocols [26] and video recordings [26, 88] to study pathologist behavior during interpretation of biopsy slides.

Treanor *et al.* [110] used software to track pathologists, two experts and two trainees, while they interpreted oesophagus biopsies with a spectrum of diagnoses from benign to malignant. They asked each participant to provide a region of interest and a diagnosis for each case. They included only the cases where expert agreed on diagnosis and ROI in their analysis and considered expert agreement as ground truth. They defined 'significant pauses' as a pause longer than 1 second at a magnification of 10x or higher. Finally, they *manually* analyzed the viewing logs by producing significant pauses and heatmaps showing how much time a participant spent on each pixel and categorized the diagnostic errors in two ways: 1) as undercalls (underdiagnoses) or overcalls (overdiagnoses), 2) as identification failures or misinterpretations. They concluded that undercalls were more common than overcalls, and all of the diagnostic errors were caused by misinterpretation, although a few of them also included failure to identify features on the slide. Although Treanor *et al.*'s study is relevant to ours in terms of tracking data analysis, their conclusions are limited by the small sample size (of cases and participants), and their analysis lacks the subjectivity and power of an automated system.

Mello-Thoms *et al.* [70] conducted a study on skin biopsies to investigate the effect of training on the diagnostic process. They experimented on two groups of pathologists, one had undergone dermatopathology training and the other had not, by tracking their interactions with a digital slide and recording their findings. They used simple measures of 1) area explored on digital slide in low, medium and high magnification, 2) time to identify diagnostically important regions, and 3) time to interpret diagnostically important regions. They compared two groups and found out that the pathologists with the dermatopathology training correctly diagnosed significantly more slides than the other group. They discovered two types of slide exploration strategy: a focused and efficient search -associated with correct final diagnoses- and a more dispersed, time-consuming strategy -associated with incorrect final diagnoses-. This study is important as an attempt to design a *representation* of the search patterns of pathologists, but the proposed representation and the features extracted are very simple to use in a complex analysis. They do not use temporal information about the image regions visited or any image features at all.

8.2 Quantifying Viewing Patterns

A viewport scene is a rectangular part of the image that is visible on the pathologists computer monitor at any time during an interpretation. The time spent on each viewport scene was calculated using logged timestamps. If an entry exceeded a total duration of five minutes, it was excluded under the assumption that the pathologist was not actively interpreting during that time. From the tracking logs, several variables were calculated to characterize the viewing behaviors of each participant, as described below.

8.2.1 Average zoom level, maximum zoom level, and zoom level variance

The web-based viewer allowed zoom levels from 1x to 60x. For each interpretation, viewport tracking logs provided a variable number of zoom level values depending on pathologists interpretive behavior; for this reason, summary statistics were used to describe zoom level behavior during each interpretation. Average and maximum zoom levels, as well as zoom level variance, were calculated for each interpretation. For each interpretation, we calculated the average zoom level

by summing the zoom level values of all viewport scenes and dividing by the number of viewport scenes. Similarly, we calculated the maximum zoom level of each interpretation and the standard deviation of the zoom level variable as the zoom level variance.

8.2.2 Scanning percentage

We quantified scanning behavior by calculating the percentage of log entries associated with panning behavior (i.e., changing viewport scene coordinates) in each interpretation. Unlike average zoom level, maximum zoom level and zoom level variance, scanning percentage considers the changes of zoom level in consecutive log entries, regardless of the zoom level itself. In other words, scanning percentage quantifies a behavior that can manifest at different zoom levels. Scanning percentage approaches 100% when the pathologist pans across different areas of the digital image at a constant zoom, and it approaches 0% when zooming in and out at different locations, with less panning or infrequent but long distance pans at a low zoom magnification. For analysis, the scanning percentages were grouped into four categorical variables (0-20%, 20-40%, 40-60%, 60-80%, 80-100%).

8.3 Statistical Analysis

To assess how pathologist demography influenced interpretive strategy, we modeled our data using repeated-measures regressions, implementing the generalized estimating equation (GEE) approach. The model included 10 categorical predictors (factors), as detailed in Table 8.1. The model used scanning percentage as a linear dependent variable (outcome).

To assess how case order within each set of 60 cases influenced viewing behaviors, we again modeled our data using repeated-measures regressions, implementing the GEE approach. We implemented two models, both including interpretation order as the continuous predictor. We used a linear dependent variable (outcome) for both models: scanning percentage for the first model and total interpretation time per case for the second model.

To assess how interpretive strategy influenced diagnostic outcome, we conducted four separate repeated-measures analyses of variance (ANOVA) with four variables that describe the interpreta-

tive behaviors. Each model included one of four continuous variables (average zoom, maximum zoom level, zoom level variance, or scanning percentage) and one of three categorical dependent variables for diagnostic outcome (over-interpretation compared to the expert consensus diagnosis, concordance with the expert consensus diagnosis, and under-interpretation compared to the expert consensus diagnosis). To assess the effect of interpretative behaviors on diagnostic efficiency, we used a repeated-measures ANOVA with a continuous dependent variable (time) and one of four independent categorical variables (scanning percentage: 0-20%, 20-40%, 40-60%, 60-80% or 80-100%).

8.4 Two Distinct Patterns: Scanners and Drillers

Viewport tracking data from 87 pathologists were analyzed, producing a total of 5,220 interpretations. Tracking logs were visualized and analyzed to summarize the interpretive strategy of each pathologist. Figure 8.1 contrasts visualizations representing two different pathologists. The pathologist represented on the left, a scanner, chose a consistent zoom level and systematically panned to investigate the whole image. The scanner pathologist used the same zoom level on the majority of their cases. In contrast, the pathologist represented on the right, a driller, zoomed out periodically, selected a new area to view, then zoomed in again. The driller pathologist zoomed in and out on different regions throughout their interpretations. It could be argued that the driller scanned the image with eye movements (rather than screen pans) at a lower resolution to determine areas for drilling. Some of the scanning versus drilling strategies may reflect the pathologists comfort level when scanning with eye movements at lower magnifications. The scanning percentage for the visualization on the left is close to 100%, while it is closer to 0% for the visualization on the right.

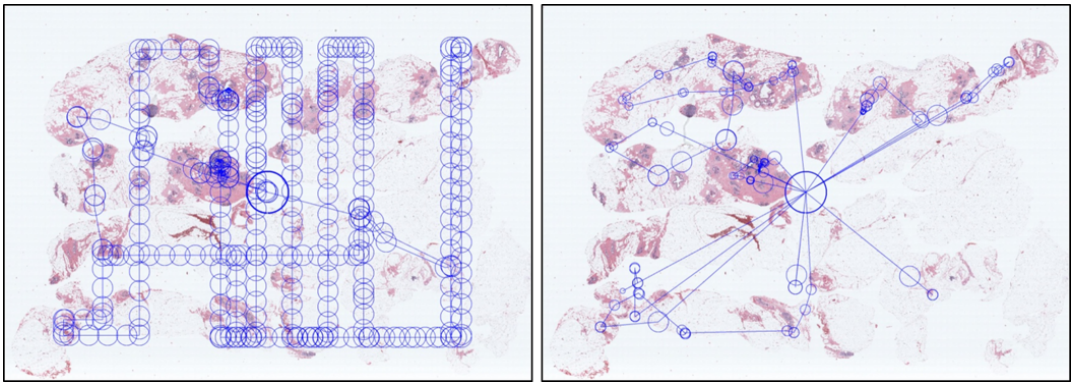


Figure 8.1: Visualization of viewport tracking logs a scanner (left) and a driller (right) on the same image. Each participant starts at the center of the image with a zoom level of 1x. The rings indicate the center of each viewport, the size of the rings indicate the zoom level (the larger the ring, the lower the zoom level), the thickness of the rings indicate the time spent at that viewport, and the lines connect consecutive viewports.

8.5 Pathologist Demographics and Scanning Behavior

Characteristics of the participating pathologists are shown in Table 8.1 along with the scanning percentage for each characteristic. Overall, pathologists tended to show scanning percentages exceeding 50% ($\mu = 74\%, \sigma = 16\%$), demonstrating a disproportionate trend toward scanning rather than drilling. This pattern was confirmed with a one-sample t-test comparing to 50%, $t(86) = 13.53, p < .001$. However, this pattern also varied significantly as a function of certain pathologist demographics.

The GEE model goodness-of-fit was 1140.98 (QIC), with three significant main effects. First, age positively predicted increasing scanning percentage ($\chi^2 = 8.25, p < .05$), with higher age groups showing increasingly higher scanning percentages. Second, there were higher scanning percentages among female versus male pathologists ($\chi^2 = 4.44, p < .05$). Finally, facility group size negatively predicted scanning percentage ($\chi^2 = 5.48, p < .05$), with pathologists working in larger facility groups showing lower scanning percentages. No other patterns reached traditional ($\alpha = .05$) significance levels.

Table 8.1: Characteristics and average scanning percentages of pathologists (N=87) ©2017, Journal of Digital Imaging

Variable	*# Pathologists*		*Average Scanning Percentage*	*P-value*	*Wald Chi-Square*
Age at survey (years)					
30-39	10	(11%)	69%	0.041	8.251
40-49	25	(29%)	77%		
50-59	36	(41%)	75%		
60+	16	(18%)	70%		
Gender					
Male	57	(66%)	70%	0.035	4.439
Female	30	(34%)	82%		
Affiliation with academic medical center					
Yes	19	(22%)	77%	0.642	0.216
No	68	(78%)	73%		
Facility size					
$< 10 pathologists$	55	(63%)	76%	0.019	5.848
$\geq 10 pathologists$	32	(37%)	69%		
Fellowship training in surgical or breast pathology					
No	41	(47%)	75%	0.076	3.141
Yes	46	(53%)	73%		
Do your colleagues consider you an expert in breast pathology?					
No	70	(80%)	73%	0.103	2.666
Yes	17	(20%)	79%		
Breast pathology experience (years)					
< 20	65	(75%)	76%	0.073	3.210
≥ 20	22	(25%)	68%		
Number of breast cases per week					
< 5	19	(22%)	73%	0.490	1.426
$5 - 9$	36	(41%)	75%		
≥ 10	32	(37%)	72%		
Marked an ROI					
Yes	44	(51%)	71%	0.565	0.330
No	43	(49%)	77%		
How confident are you in your assessments of breast cases?					
1 (very confident)	13	(15%)	67%	0.100	7.783
2	36	(41%)	75%		
3	32	(21%)	75%		
4	8	(9%)	77%		
5 (not confident at all)	2	(2%)	83%		

8.6 Interpretation Order and Scanning Behavior

The GEE model showed a significant negative relationship between case position and scanning percentage ($\chi^2 = 16.01, p < .001$), with scanning percentage decreasing over the course of the 60 cases (see Figure 8.2). The total time spent on an interpretation of each case also decreased on average with interpretation order. The participants interpreted later cases in less time compared to earlier cases, ($\chi^2 = 67.36, p < .001$).

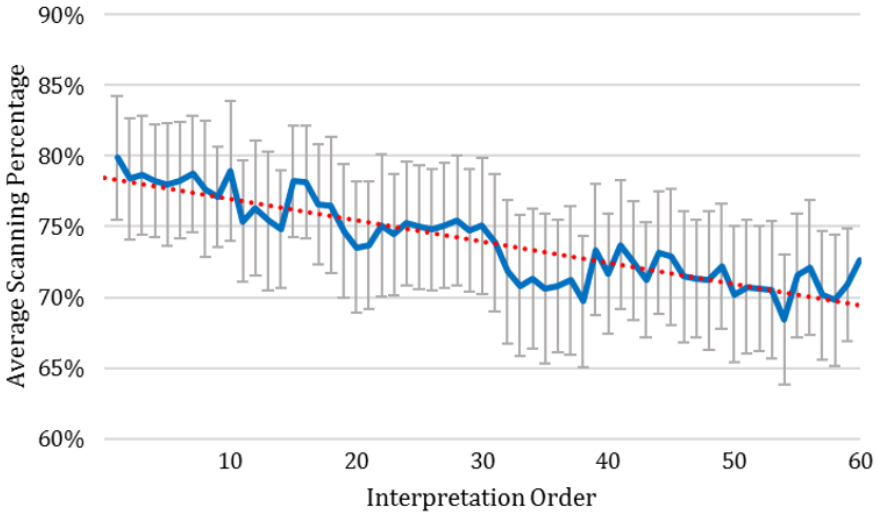


Figure 8.2: Average scanning percentage of 87 pathologists during the interpretation of 60 test cases. The order of the 60 cases was randomized for each pathologist so that n-th case included a random sampling of cases from all diagnostic categories. ©2017, Journal of Digital Imaging

8.7 Scanning Behavior on Diagnostic Concordance

The mean values of the average zoom level, maximum zoom level, zoom level variance, and scanning percentage variables for interpretations are shown by expert consensus diagnosis and concordance with expert consensus diagnosis in Table 8.3. Table 8.2 provides detailed results of ANOVA tests.

Over-interpretation was associated with increased drilling (average zoom level, maximum zoom level, and zoom level variance). Average zoom level, maximum zoom level, and zoom level vari-

Table 8.2: Detailed statistical outcomes of repeated-measures ANOVA for assessing how interpretative behavior affects diagnostic outcome. ©2017, Journal of Digital Imaging

		factor:diagnostic concordance					*error*		
		sum of squares	*df*	*mean square*	*F*	*p-val*	*sum of squares*	*df*	*mean square*
all	*Average Zoom Level*	160.71	2	80.36	33.30	≤ 0.01	415.01	172	2.41
	Maximum Zoom Level	1999.17	2	999.59	44.02	≤ 0.01	3905.50	172	22.71
	Zoom Level Variance	143.27	2	71.64	40.07	≤ 0.01	307.51	172	1.79
	Scanning Percentage	0.003	2	0.002	0.56	0.57	0.47	172	0.003
benign	*Average Zoom Level*	415.12	1	415.12	72.50	≤ 0.01	458.04	80	5.73
	Maximum Zoom Level	3396.43	1	3396.43	55.56	≤ 0.01	4890.37	80	61.13
	Zoom Level Variance	305.68	1	305.68	53.41	≤ 0.01	457.82	80	5.72
	Scanning Percentage	0.007	1	0.007	1.19	0.28	0.48	80	0.006
atypia	*Average Zoom Level*	178.00	2	90.00	17.57	≤ 0.01	788.90	154	5.12
	Maximum Zoom Level	2661.63	2	1330.81	31.94	≤ 0.01	6417.23	154	41.67
	Zoom Level Variance	173.74	2	86.87	20.33	≤ 0.01	658.09	154	4.27
	Scanning Percentage	0.01	2	0.007	1.30	0.28	0.78	154	0.005
DCIS	*Average Zoom Level*	114.04	2	57.02	9.38	≤ 0.01	364.78	60	6.08
	Maximum Zoom Level	678.12	2	339.06	8.20	0.001	2482.22	60	41.37
	Zoom Level Variance	47.34	2	23.67	6.33	0.003	224.53	60	3.74
	Scanning Percentage	0.016	2	0.008	1.02	0.37	0.47	60	0.01
invasive	*Average Zoom Level*	10.16	1	10.16	1.16	0.31	69.80	8	8.73
	Maximum Zoom Level	7.35	1	7.35	0.08	0.79	777.52	8	97.19
	Zoom Level Variance	0.80	1	0.80	0.10	0.76	65.46	8	8.18
	Scanning Percentage	0.002	1	0.002	0.22	0.65	0.07	8	0.008

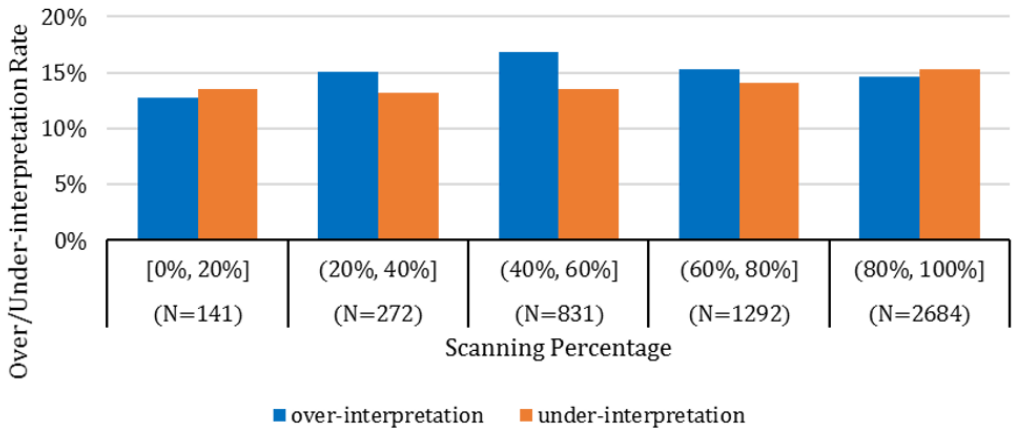


Figure 8.3: Over- and under- interpretation rates in five categories of scanning percentage. ©2017, Journal of Digital Imaging

ance were higher than the expert consensus diagnosis for over-interpretations and were lower than the expert consensus diagnosis for under-interpretations. The trend was replicated in benign, atypia, and invasive cases. For DCIS cases, both over-interpretation and under-interpretation were associated with higher zoom values. All associations except those for invasive cases were statistically significant.

No association was noted between scanning percentage and accuracy. Figure 8.3 shows the average over-interpretation and under-interpretation rates within different scanning percentage groups.

Table 8.3: Zoom and scanning variables by concordance with expert consensus diagnosis. ©2017, Journal of Digital Imaging

Consensus Diagnosis	*Diagnostic Concordance*	*Number of Interpretations*	*Average Zoom Level*	*p-val*	*Maximum Zoom Level*	*p-val*	*Zoom Level Variance*	*p-val*	*Scanning Percentage*	*p-val*
	under	760	7.89		24.93		6.22		75%	
all	*agree*	3672	8.86	≤ 0.01	27.29	≤ 0.01	6.98	≤ 0.01	74%	0.574
	over	788	9.94		31.87		8.10		73%	
	under	-	-		-		-		75%	
benign	*agree*	933	6.64	≤ 0.01	22.10	≤ 0.01	5.28	≤ 0.01	75%	0.278
	over	348	9.71		31.49		7.98		72%	
	under	492	7.39		23.21		5.78		76%	
atypia	*agree*	882	8.88	≤ 0.01	27.51	≤ 0.01	7.03	≤ 0.01	72%	0.276
	over	384	10.00		31.79		8.13		73%	
	under	259	8.74		28.14		7.03		73%	
DCIS	*agree*	882	8.36	≤ 0.01	27.66	0.001	6.88	0.003	74%	0.365
	over	384	10.95		34.82		8.64		78%	
	under	9	10.40		26.67		7.73		74%	
invasive	*agree*	471	14.72	0.312	36.10	0.79	6.88	0.763	75%	0.652
	over	-	-		-		-		-	

8.8 Scanning Behavior and Diagnostic Efficiency

Efficiency to arrive at an accurate diagnosis was negatively predicted by the extent to which pathologists followed a scanning strategy; in other words, higher scanning percentage was associated with lower efficiency. A repeated-measures ANOVA revealed a main effect of scanning percentage category, $F(4, 52) = 6.72, p < .001$, demonstrating significantly higher case review times as a function of increased scanning percentage. This pattern is depicted in Figure 8.4. Follow-up paired t-tests demonstrated significant differences between all pairwise category comparisons, with the exception of the first (0-20%) versus second (20-40%) categories, and fourth (60-80%) versus fifth (80-100%) categories. In contrast, rates of diagnostic concordance with the expert consensus diagnosis showed no significant difference across scanning percentage groups.

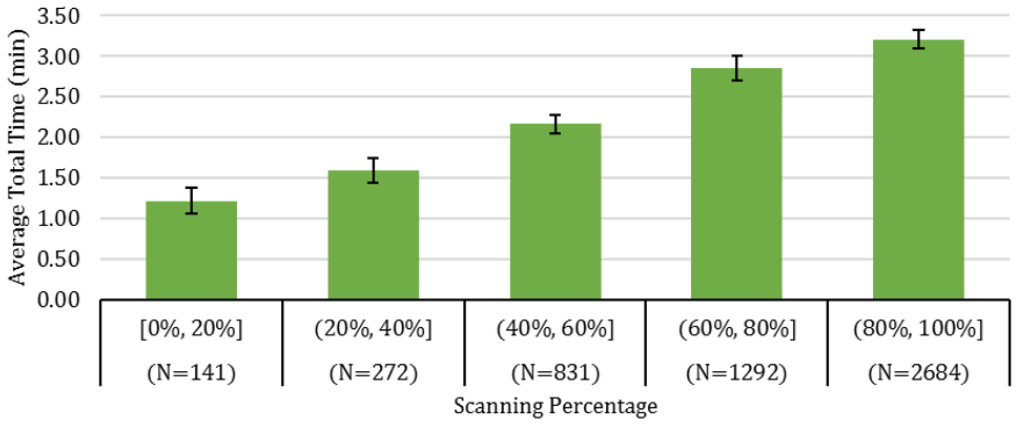


Figure 8.4: Average total time of interpretation in five categories of scanning percentage. ©2017, Journal of Digital Imaging

8.9 Discussion

The field of pathology has begun adopting the digital WSI format as it offers great potential for teaching [15, 62, 96, 121] and research [51], as well as archival purposes [42] and gathering second opinions [5, 85]. To better understand the visual search patterns used in breast pathology, 87 pathologists across the U.S. interpreted 60 digital WSIs of breast biopsies representing a range of

diagnostic categories, amounting to 5,220 individual independent interpretations for analysis.

A web-based viewer tracked and recorded the interpretive behaviors of pathologists as they viewed each digital WSI. The viewer provided pathologists with two possible actions: zooming and panning. Zooming in to an area allowed pathologists to examine cytological, cellular, and nuclear structural details, thereby revealing those that are not as visible to the human eye at lower magnification, but also limiting the portion of the whole slide image viewable on the screen. The panning action allowed pathologists to view neighboring areas of the whole slide image that were not viewable on the screen at higher magnifications.

Combinations of both actions were used by all pathologists to interpret the digital WSI, but interpretive patterns emerged when we analyzed the tracking logs. Some participants, namely drillers, preferred to use panning actions sparingly to examine small areas around a focal point, and then continued their interpretation by zooming out to a lower magnification. Participants at the other end of the spectrum, namely scanners, used panning actions to systematically explore the complete image at a constant magnification. As scanning and drilling are complementary strategies, participants used a combination of both, but in different ratios. We wanted to explore potential explanations for the two interpretative strategies through their correlation with diagnostic accuracy and efficiency, as well as determining if these patterns change over time.

A number of pathologist demographic characteristics were associated with changes in scanning percentage, including age, gender, and facility size. Higher age was positively correlated with increased scanning percentage, females scanned more than males, and pathologists from smaller facility sizes had higher scanning percentages. It may have been the case that younger participants had relatively more prior experience with similar computer interfaces or image manipulation tools (e.g., mapping software, digital slide viewers, image editing software), thereby making them more comfortable with image drilling behaviors [32]. Although not a statistically significant trend, pathologists with higher scanning percentages also reported lower baseline confidence in their breast pathology skills. This finding suggests that increased scanning may be related to personality-level (e.g., neuroticism [79]) and/or situation-level (e.g., anxiety [117]) factors.

The scanning percentage and total time per slide decreased as pathologists gained experience

throughout the set of 60 cases. This suggests a learning curve where participants who started with a scanning-based strategy adopted a more hybrid approach of scanning and drilling as they interpreted through the digital images. This learning curve may be due to prior inexperience with digital slides and computer-based viewing systems that pathologists began to overcome through their experience in this study. Previous research shows a learning curve for interpreting mammograms before and after residency, suggesting a correlation between interpretive behavior and experience [74].

We also noted a pattern of over-interpretation at higher zoom levels. For all diagnostic categories except invasive cancer, the cases that were over-interpreted based on the expert consensus diagnosis had higher values of average zoom level, maximum zoom level, and zoom level variance. This relationship aligns with some research in the cognitive science and visual search literature. Specifically, when observers repeatedly examine a visual scene in detail, the probability of making an erroneous guess increases [22]. These inaccurate interpretations likely result from a failed match between perceived image features and stored histopathological features in their memory.

In order to analyze the association of scanning with accuracy and total interpretation time, we divided participants and interpretations into five groups based on scanning percentage. Interpretation time increased in each increasing scanning percentage group. However, over-interpretation and under-interpretation rates were similar for all scanning percentage groups. Scanning was found to be a less efficient strategy for diagnostic interpretation, and the results with the learning curve indicated that pathologists adopted a more balanced and efficient strategy as they progressed through the set of 60 cases. Recent research on volumetric lung images revealed that radiologists adopt distinct visual search strategies during interpretation[31]. Though this earlier research used eye-tracking to monitor and interpret visual search patterns, our findings suggests that similar distinctions can be ascertained by recording zooming and panning behaviors. We expect this is specifically the case with 2D digital pathology images. Indeed, these images require pathologists to zoom in and out dramatically in order to magnify breast tissue and reveal specific structural and cellular features. This process results in high-density zooming and panning data, which is likely uncharacteristic of viewing behavior with narrow slices of volumetric images. The unique characteristics of

these breast biopsy digital WSIs may explain why our data did not suggest any influence of visual search strategy on diagnostic accuracy, unlike earlier research with volumetric lung radiographs [31, 116]. Of course, when attempting to identify specific structural or cellular features that were viewed or neglected during the interpretive process, eye-tracking is an invaluable technique.

Limitations and Strengths

This study was limited to one slide per case, which does not reflect actual clinical practicea factor that may influence diagnostic accuracy but does not preclude evaluation of interpretive strategies. However, the one-slide-per-case study design reduced the workload of participants and allowed them to interpret more images representing a variety of tissue characteristics. This study also diverged from clinical practice in the distribution of diagnostic categories among the cases participants interpreted. Atypia and DCIS cases were oversampled compared to actual clinical prevalence, with the purpose of better understanding the interpretation of these diagnostically difficult non-invasive cases. Previous research shows that atypia and DCIS cases are more likely to be overinterpreted or misinterpreted, so it is crucial to understand interpretive behaviors on these diagnoses [34]. A possible limitation was the participants prior inexperience with the digital format or digital viewers. Although the field of pathology has begun to incorporate digital WSI, most U.S. pathologists are still inexperienced with software for digital WSI interpretation, making it difficult to dissociate the relative contributions of experience with the digital format versus expertise in breast pathology on drilling versus scanning. Similarly, some variation between pathologists may be attributable to participants using their own computer monitors; it is therefore possible that identical monitors may standardize the pathologists experience viewing digital WSI. However, identical monitors do not reflect actual clinical practice, where monitors vary at the level of the practice and, often, between pathologists at the same facility.

Limitations aside, this study is a timely and unique investigation of pathologists interpretive strategies with digital media. Strengths of this study include the large sample size of breast biopsy cases representing a full spectrum of diagnostic categories from benign and atypia to DCIS and invasive cancer. Another strength is the large number of practicing pathologists from across the

U.S. The use of a web-based viewer allowed participants to use their own computers in their own time, which is as close to the real-world practice of digital pathology as possible. Furthermore, the order of the 60 cases was random for each participant, which allowed us to see a learning curve for the digital slide viewer without case biases attributable to interpretive difficulty or severity of diagnosis.

8.10 Summary

Digital slides are increasingly being used in research, education, archiving and obtaining second opinions remotely. To understand the diagnostic decision making process and sources of interpretation errors, we studied a group of 87 pathologists interpretative behaviors on a set of 240 digital whole slide images of breast biopsy specimens that spanned the diagnostic categories of benign, atypia, ductal carcinoma in situ, and invasive cancer. A web-based virtual viewer recorded pathologist behaviors during their interpretations of a subset of 60 randomly selected and randomly ordered slides. Two distinct strategies were common during interpretation: scanning and drilling. Scanning involves panning at a constant zoom level, while zooming in and out at various locations is drilling. We used the location of the screen on the whole slide image, time stamp and zoom level data to calculate four variables to further define the interpretative behavior: average zoom level, maximum zoom level, zoom level variance, and scanning percentage. Through statistical analysis, we found that females scanned more than males and age was positively correlated with scanning percentage, while the facility size was negatively correlated. We found that as each pathologist sequentially interpreted 60 cases, the scanning percentage and total interpretation time per slide decreased, and these two variables were positively correlated. However, the percentage of time spent scanning or drilling was not predictive of diagnostic accuracy. Finally, we identified a trend in which increasing average zoom level, maximum zoom level, and zoom variance correlated with over-interpretation of atypia and DCIS. In summary, the interpretation strategy pathologists use can be affected by their backgrounds and experience. Furthermore, the interpretation strategy can affect the diagnostic outcome and efficiency.

Chapter 9

CONCLUSIONS

The purpose of this work was to understand the diagnostic errors by investigating the viewing behavior of the pathologists and the image characteristics of the digital slides of breast biopsies. We developed novel analysis methods for the viewing behavior and novel image features for the pre-invasive and invasive lesions of the breast. We used the digiPATH dataset that contains 240 whole slide images of breast biopsies with four diagnostic categories: benign, atypia, DCIS and invasive cancer. The digiPATH dataset also provides us with the unique opportunity to study the interpretive process using the viewport tracking logs of 87 participating and three expert pathologists.

We first tackled the problem of identifying diagnostically relevant regions on whole slide images (Chapter 4). Using the tracking logs from three experts, we developed a novel method to extract the regions they focused on based on their zooming and panning actions. Then, we developed a visual bag-of-words model based on color and texture features of the image to automatically detect these regions on unseen images. We achieved a 74% accuracy in the detection of the regions of interest on whole slide images.

Then in Chapters 5 and 6, we developed image features to describe the ROIs and validated them in diagnostic classification experiments. We used a tissue labeling scheme to summarize the tissue composition of the images: the complete eight labels include the background, benign epithelium, malignant epithelium, normal stroma, desmoplastic stroma, secretion, blood and necrosis. Using a subset (N=58) of the digiPATH ROIs, we collected ground truth pixel labels from a pathologist. Then, we developed two models, a superpixel-based model and a CNN-based model, for the semantic segmentation of the ROIs. While the superpixel-based model achieved an average F_1 score of .35 for the five labels and .34 for the eight labels, the CNN-based model achieved an F_1 score of .39 for the five labels and .25 for the eight labels. Using the label images produced by the semantic

segmentations, we developed two features the superpixel-label frequency and co-occurrence histograms and the structure feature. We achieved 0.91 classification accuracy for the invasive cancer, 0.72 classification accuracy for the benign cases and 0.88 classification accuracy, which is higher than the participant's performance, for the atypia and DCIS cases.

Finally, in Chapters 7 and 8, we analyzed the interpretation patterns of the pathologists. We found that the pathologists identifying an ROI that overlaps with the consensus ROI determined by the experts have higher concordance with the consensus diagnosis. In other words, there is a considerable amount of diagnostic error caused by misidentifying the ROI. We also analyzed the viewport tracking logs of the pathologists and identified two behavior patterns: scanning and drilling. Although there is no significant difference between the diagnostic concordance of scanners and drillers, drillers are much more efficient in their diagnostic decision making. Furthermore, we found a learning curve towards drilling in the course of 60 cases each pathologist interpreted.

9.1 Contributions

We developed a novel way of analyzing tracking logs of the pathologists the digiPATH project offered. We defined three key actions that correspond to the regions pathologists paid attention to. We also developed a scanning percentage measure that quantifies the scanning behavior. To our best knowledge, there is no formal methodology to analyze the tracking logs of pathologists on whole slide images. We attempted to automate the process and develop objective measures for the interpretation patterns.

The structure feature is a novel way of describing the structural changes that occur in the breast ducts during the course of cancer. We proposed a semantic segmentation of the biopsy images using tissue labels to produce an abstract representation of the objects in the image. We developed the structure feature specifically for the pre-invasive lesions of the breast. The histopathological image analysis literature is full of methods classifying invasive cancer images but our structure feature is one of the first attempts to describe the pre-invasive lesions.

9.2 Future Work

Cellular Features

Cancer changes the mechanism of a cell and affects its appearance. In ductal breast cancer, the appearance of the epithelial cells changes the diagnosis. The degree of nuclear atypia and the irregularities in the organization of cells are evaluated by the pathologist during the interpretation of the biopsy.

In our work, we focused on the structures, or ducts, rather than the epithelial cells that make up the structures. We encoded cellular malignancy in our tissue label segmentation (Chapter 5) by using two labels for the epithelial tissue: benign and malignant. In the diagnostic classification experiments, we tested our mid-level and high-level features with the eight tissue labels and the five tissue labels (merging benign and malignant epithelium labels). The features with the eight tissue labels produced better classification accuracies than those with the five tissue labels, especially in the atypia vs. DCIS and the invasive vs. non-invasive tasks. This result shows the importance of the cellular level differentiation. Furthermore, all of our features struggled with the benign vs. atypia-DCIS task. One possible explanation for this is that the differentiation between benign and atypia is mostly based on the level of cellular atypia. With our two labels for the epithelium, we could not capture the wide range of atypia that is present in the benign and atypical proliferations. Adding cellular features may provide a fine-detailed description of the cellular changes and improve the classification of the benign cases.

In addition to simple statistics of the cells (like size, shape and texture), the organization of the cells can be encoded using different graph structures that use the cells as the vertices. By extracting simple features from the graphs constructed over the cells, it may be possible to quantify the regularity of the organization and orientation of the cells.

Image Characteristics vs. Viewing Behavior

In this thesis, we described methodologies for analyzing viewing behavior for region of interest detection (Chapter 4) and the viewing pattern quantification (Chapter 8). We also studied image

features for the automated detection of the regions of interest (4) and the diagnostic classification of the regions of interest (Chapter 6). We analyzed the effect of viewing behavior (Chapter 8) and region of interest identification (7) on the diagnostic accuracy. One missing analysis is the relationship between the image characteristics, viewing behavior and diagnostic accuracy.

The image features we developed are powerful enough to differentiate between different diagnostic categories. They can be used to describe any image portion the pathologists looked at. Combining the methodologies we developed for viewing behavior analysis with the image features, one can explore which image features attract the pathologist's attention and more importantly which ones lead to an incorrect diagnosis. This piece of the puzzle is left as a future work.

BIBLIOGRAPHY

[1] FDA allows marketing of first whole slide imaging system for digital pathology. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm552742.htm, .

[2] The Cancer Genome Atlas. http://cancergenome.nih.gov, .

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012.

[4] S. Akbar, L. Jordan, A. M. Thompson, and S. J. McKenna. Tumor localization in tissue microarrays using rotation invariant superpixel pyramids. In *2015 IEEE 12th International Symposium on Biomedical Imaging*, pages 1292–1295, Apr 2015.

[5] S. Al-Janabi, A. Huisman, and P. J. Van Diest. Digital pathology: current status and future perspectives. *Histopathology*, 61(1):1–9, 2012.

[6] K. H. Allison, L. M. Reisch, P. A. Carney, D. L. Weaver, S. J. Schnitt, F. P. O'Malley, B. M. Geller, and J. G. Elmore. Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel. *Histopathology*, 65(2):240–251, Aug 2014.

[7] I. Arganda-Carreras, S. C. Turaga, D. R. Berger, D. Cirean, A. Giusti, L. M. Gambardella, J. Schmidhuber, D. Laptev, S. Dwivedi, J. M. Buhmann, T. Liu, M. Seyedhosseini, T. Tasdizen, L. Kamentsky, R. Burget, V. Uher, X. Tan, C. Sun, T. D. Pham, E. Bas, M. G. Uzunbas, A. Cardona, J. Schindelin, and H. S. Seung. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in Neuroanatomy*, 2015.

[8] C. Bahlmann, A. Patel, J. Johnson, J. Ni, A. Chekkoury, P. Khurd, A. Kamen, L. Grady, E. Krupinski, A. Graham, and R. Weinstein. Automated detection of diagnostically relevant regions in H&E stained digital pathology slides. In *Proceedings of SPIE, Medical Imaging*, page 831504. International Society for Optics and Photonics, Feb 2012.

[9] W. E. Barlow, E. White, R. Ballard-Barbash, P. M. Vacek, L. Titus-Ernstoff, P. A. Carney, J. A. Tice, D. S. M. Buist, B. M. Geller, R. Rosenberg, B. C. Yankaskas, and K. Kerlikowske. Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography. *Journal of the National Cancer Institute*, 98(17):1204–1214, Sep 2006.

[10] A. Basavanhally, S. Ganesan, M. Feldman, N. Shih, C. Mies, J. Tomaszewski, and A. Madabhushi. Multi-Field-of-View Framework for Distinguishing Tumor Grade in ER+ Breast Cancer From Entire Histopathology Slides. *IEEE Transactions on Biomedical Engineering*, 60(8):2089–2099, Aug 2013.

[11] C. A. Beam, E. F. Conant, and E. A. Sickles. Correlation of radiologist rank as a measure of skill in screening and diagnostic interpretation of mammograms. *Radiology*, 238(2): 446–453, Feb 2006.

[12] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108): 108ra113–108ra113, 2011.

[13] B. E. Bejnordi, G. Litjens, M. Hermsen, N. Karssemeijer, and J. A. W. M. van der Laak. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, volume 9420, 2015.

[14] D. Bombari, B. Mora, S. C. Schaefer, F. W. Mast, and H. A. Lehr. What Was I Thinking?

Eye-Tracking Experiments Underscore the Bias that Architecture Exerts on Nuclear Grading in Prostate Cancer (Cognitive Bias in Diagnostic Pathology). *PLoS ONE*, 7(5):e38023, 2012.

[15] L. A. Bruch, B. R. De Young, C. D. Kreiter, T. H. Haugen, T. C. Leaven, and F. R. Dee. Competency assessment of residents in surgical pathology using virtual microscopy. *Human Pathology*, 40(8):1122–1128, Aug 2009.

[16] T. T. Brunyé, P. A. Carney, K. H. Allison, L. G. Shapiro, D. L. Weaver, and J. G. Elmore. Eye Movements as an Index of Pathologist Visual Expertise: A Pilot Study. *PLoS ONE*, 9 (8):e103447, Aug 2014.

[17] T. T. Brunyé, M. D. Eddy, E. Mercan, K. H. Allison, D. L. Weaver, and J. G. Elmore. Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation. *BMC Medical Informatics and Decision Making*, 16(1):1–8, 2016.

[18] T. T. Brunyé, E. Mercan, D. L. Weaver, and J. G. Elmore. Accuracy is in the Eyes of the Pathologist: The Visual Interpretive Process and Diagnostic Accuracy with Digital Whole Slide Images. *Journal of Biomedical Informatics*, 66:171–179, 2017.

[19] H. Chang, G. V. Fontenay, J. Han, G. Cong, F. L. Baehner, J. W. Gray, P. T. Spellman, and B. Parvin. Morphometic analysis of TCGA glioblastoma multiforme. *BMC Bioinformatics*, 12(1):484, 2011.

[20] A. Chekkoury, P. Khurd, J. Ni, C. Bahlmann, A. Kamen, A. Patel, L. Grady, M. Singh, M. Groher, N. Navab, E. Krupinski, J. Johnson, A. Graham, and R. Weinstein. Automated malignancy detection in breast histopathological images. In *SPIE Medical Imaging*, volume 8315, page 831515. International Society for Optics and Photonics, Feb 2012.

[21] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng. DCAN: Deep contour-aware networks for object instance segmentation from histology images. In *Medical Image Analysis*, volume 36, pages 135–146, Feb 2017.

[22] M. M. Chun and J. M. Wolfe. Just Say No: How Are Visual Searches Terminated When There Is No Target Present? *Cognitive Psychology*, 30(1):39–78, Feb 1996.

[23] D. C. Cirean, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *Proc Medical Image Computing Computer Assisted Intervenction (MICCAI)*, pages 411–418. 2013.

[24] L. A. D. Cooper, J. Kong, D. A. Gutman, W. D. Dunn, M. Nalisnik, and D. J. Brat. Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images. *Lab Invest*, 95(4):366–376, 2015.

[25] C. Cortes and V. Vapnik. Support-vector networks, 1995.

[26] R. S. Crowley, G. J. Naus, J. Stewart, and C. P. Friedman. Development of visual diagnostic expertise in pathology: An information-processing study. *Journal of the American Medical Informatics Association*, 10(1):39–51, 2003.

[27] M. D. DiFranco, G. O'Hurley, E. W. Kay, R. W. G. Watson, and P. Cunningham. Ensemble based system for whole-slide prostate cancer probability mapping using color texture features. *Computerized Medical Imaging and Graphics*, 35(7):629–45, 2011.

[28] F. Dong, H. Irshad, E. Y. Oh, M. F. Lerwill, E. F. Brachtel, N. C. Jones, N. W. Knoblauch, L. Montaser-Kouhsari, N. B. Johnson, L. K. F. Rao, B. Faulkner-Jones, D. C. Wilbur, S. J. Schnitt, and A. H. Beck. Computational pathology to discriminate benign from malignant intraductal proliferations of the breast. *PLoS ONE*, 9(12):e114885, 2014.

[29] 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 *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 496–499, May 2008.

[30] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi. A Boosted Bayesian Multiresolution Classifier for Prostate Cancer Detection From Digitized Needle Biopsies. *IEEE Transactions on Biomedical Engineering*, 59(5):1205–1218, May 2012.

[31] T. Drew, M. L.-H. Vo, A. Olwal, F. Jacobson, S. E. Seltzer, and J. M. Wolfe. Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, 13(10):3–3, Aug 2013.

[32] S. M. Elias, W. L. Smith, and C. E. Barney. Age as a moderator of attitude towards technology in the workplace: work motivation and overall job satisfaction. *Behaviour & Information Technology*, 31(5):453–467, May 2012.

[33] J. G. Elmore. Solving the problem of overdiagnosis. *New England Journal of Medicine*, 375(15):1483–1486, 2016.

[34] J. G. Elmore, G. M. Longton, P. A. Carney, B. M. Geller, T. Onega, A. N. A. Tosteson, H. D. Nelson, M. S. Pepe, K. H. Allison, S. J. Schnitt, F. P. OMalley, and D. L. Weaver. Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens. *JAMA*, 313(11): 1122, Mar 2015.

[35] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010.

[36] A. Fakhry, T. Zeng, and S. Ji. Residual Deconvolutional Networks for Brain Electron Microscopy Image Segmentation. *IEEE Transactions on Medical Imaging*, 36(2):447–456, Feb 2017.

[37] Food and Drug Administration. Classify Your Medical Device. `http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ClassifyYourDevice/`, 2016.

[38] T. J. Fuchs, P. J. Wild, H. Moch, and J. M. Buhmann. Computational Pathology Analysis of Tissue Microarrays Predicts Survival of Renal Clear Cell Carcinoma Patients. In *Medical Image Computing and Computer-Assisted Intervention*, volume 5242, pages 1–8. 2008.

[39] L. Gorelick, O. Veksler, M. Gaed, J. A. Gomez, M. Moussa, G. Bauman, A. Fenster, and A. D. Ward. Prostate Histopathology: Learning Tissue Component Histograms for Cancer Detection and Classification. *IEEE Transactions on Medical Imaging*, 32(10):1804–1818, Oct 2013.

[40] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer. Automatic segmentation of colon glands using object-graphs. *Medical Image Analysis*, 14(1):1–12, Feb 2010.

[41] R. Gutiérrez, F. Gómez, L. Roa-Peña, and E. Romero. A supervised visual model for finding regions of interest in basal cell carcinoma images. *Diagnostic Pathology*, 6(1):26, 2011.

[42] D. A. Gutman, J. Cobb, D. Somanna, Y. Park, F. Wang, T. Kurc, J. H. Saltz, D. J. Brat, L. A. D. Cooper, and J. Kong. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *Journal of the American Medical Informatics Association*, 20(6):1091–1098, Nov 2013.

[43] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision*, volume 11-18-Dece, pages 1026–1034, Dec 2015.

[44] G. E. Hinton. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313 (5786):504–507, 2006.

[45] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz. Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, volume 2016, pages 2424–2433, 2016.

[46] N. Howlader, A. Noone, M. Krapcho, D. Miller, K. Bishop, S. Altekruse, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. Lewis, H. Chen, E. Feuer, and K. e. Cronin. SEER Cancer Statistics Review, 1975-2013, National Cancer Institute. Technical report, Apr 2016.

[47] C.-H. Huang, A. Veillard, L. Roux, N. Loménie, and D. Racoceanu. Time-efficient sparse analysis of histopathological whole slide images. *Computerized Medical Imaging and Graphics*, 35(7-8):579–591, Oct 2011.

[48] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Arxiv*, pages 1–11, 2015.

[49] H. Irshad, S. Jalali, L. Roux, D. Racoceanu, G. Naour, L. Hwee, and F. Capron. Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach. *Journal of Pathology Informatics*, 4(2):12, 2013.

[50] H. Irshad, L. Roux, and D. Racoceanu. Multi-channels statistical and morphological features based mitosis detection in breast cancer histopathology. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6091–6094, Jul 2013.

[51] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu. Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review – Current Status and Future Potential. *IEEE Reviews in Biomedical Engineering*, 7:97–114, 2014.

[52] K. Jafari-Khouzani and H. Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering*, 50(6):697–704, 2003.

[53] P. Khurd, L. Grady, A. Kamen, S. Gibbs-Strauss, E. M. Genega, and J. V. Frangioni. Network cycle features: Application to computer-aided Gleason grading of prostate cancer histopathological images. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1632–1636, Mar 2011.

[54] J. Kong, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan. Image Analysis for Automated Assessment of Grade of Neuroblastic Differentiation. In *2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings*, pages 61–64, Apr 2007. ISBN 1-4244-0671-4.

[55] 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*, 42(6):1080–1092, 2009.

[56] J. Kong, L. A. D. Cooper, F. Wang, J. Gao, G. Teodoro, L. Scarpace, T. Mikkelsen, M. J. Schniederjan, C. S. Moreno, and D. J. Saltz Joel H. Brat. Machine–Based Morphologic Analysis of Glioblastoma Using Whole–Slide Pathology Images Uncovers Clinically Relevant Molecular Correlates. *PLoS ONE*, 8(11):e81049, 2013.

[57] S. Kothari, J. H. Phan, A. N. Young, and M. D. Wang. Histological Image Feature Mining Reveals Emergent Diagnostic Properties for Renal Cancer. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 422–425, Nov 2011.

[58] S. Kothari, A. O. Osunkoya, J. H. Phan, and M. D. Wang. Biological interpretation of morphological patterns in histopathological whole-slide images. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 218–225, 2012.

[59] E. A. Krupinski and R. S. Weinstein. Changes in visual search patterns of pathology residents as they gain experience. In D. J. Manning and C. K. Abbey, editors, *Proceedings of SPIE*, volume 7966, page 79660P, Mar 2011.

[60] E. A. Krupinski, A. A. Tillack, L. Richter, J. T. Henderson, A. K. Bhattacharyya, K. M. Scott, A. R. Graham, M. R. Descour, J. R. Davis, and R. S. Weinstein. Eye-movement study and human performance using telepathology virtual slides. Implications for medical education and differences with experience. *Human Pathology*, 37(12):1543–1556, 2006.

[61] E. A. Krupinski, A. R. Graham, and R. S. Weinstein. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human Pathology*, 44(3):357–364, Mar 2013.

[62] R. K. Kumar, B. Freeman, G. M. Velan, and P. J. De Permentier. Integrating histology and histopathology teaching in practical classes using virtual slides. *The Anatomical Record Part B: The New Anatomist*, 289B(4):128–133, 2006.

[63] H. L. Kundel, C. F. Nodine, E. F. Conant, and S. P. Weinstein. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*, 242(2):396–402, 2007.

[64] H. L. Kundel, C. F. Nodine, E. A. Krupinski, and C. Mello-Thoms. Using Gaze-tracking Data and Mixture Distribution Analysis to Support a Holistic Model for the Detection of Cancers on Mammograms. *Academic Radiology*, 15(7):881–886, 2008.

[65] Y. Liu, Y. Sun, R. Broaddus, J. Liu, A. K. Sood, I. Shmulevich, and W. Zhang. Integrated analysis of gene expression and tumor nuclear image profiles associated with chemotherapy response in serous ovarian carcinoma. *PLoS One*, 7(5):e36383, 2012.

[66] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 07-12-June: 3431–3440, 2015.

[67] C. Lu and M. Mandal. Automated segment.tion and analysis of the epidermis area in skin histopathological images. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5355–5359, 2012.

[68] C. Lu, M. Mahmood, N. Jha, and M. Mandal. Automated segmentation of the melanocytes in skin histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 17 (2):284–296, 2013.

[69] F. C. Martins, I. D. Santiago, A. Trinh, J. Xian, A. Guo, K. Sayal, M. Jimenez-Linan, S. Deen, K. Driver, M. Mack, J. Aslop, P. D. Pharoah, F. Markowetz, and J. D. Brenton. Combined image and genomic analysis of high-grade serous ovarian cancer reveals PTEN loss as a common driver event and prognostic classifier. *Genome Biology*, 15(12):526, 2014.

[70] C. Mello-Thoms, C. A. B. Mello, O. Medvedeva, M. Castine, E. Legowski, G. Gardner, E. Tseytlin, and R. Crowley. Perceptual analysis of the reading of dermatopathology virtual slides by pathology residents. *Archives of Pathology and Laboratory Medicine*, 136(5): 551–562, 2012.

[71] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunye, and J. G. Elmore. Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images. In *22nd International Conference on Pattern Recognition*, pages 1179–1184, 2014.

[72] E. Mercan, S. Aksoy, L. G. Shapiro, D. L. Weaver, T. T. Brunyé, and J. G. Elmore. Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images: A Comparative Study. *Journal of Digital Imaging*, 29(4):496–506, Aug 2016.

[73] E. Mercan, L. G. Shapiro, T. T. Brunye, D. L. Weaver, and J. G. Elmore. Characterizing Diagnostic Search Patterns in Digital Breast Pathology: Scanners and Drillers. *Journal of Digital Imaging*, 2017.

[74] D. L. Miglioretti, C. C. Gard, P. A. Carney, T. L. Onega, D. S. M. Buist, E. A. Sickles, K. Kerlikowske, R. D. Rosenberg, B. C. Yankaskas, B. M. Geller, and J. G. Elmore. When Radiologists Perform Best: The Learning Curve in Screening Mammogram Interpretation. *Radiology*, 253(3):632–640, Dec 2009.

[75] M. Mokhtari, M. Rezaeian, S. Gharibzadeh, and V. Malekian. Computer aided measurement of melanoma depth of invasion in microscopic images. *Micron*, 61:40–48, 2014.

[76] D. B. Nagarkar, E. Mercan, D. L. Weaver, T. T. Brunyé, P. A. Carney, M. H. Rendi, A. H.

Beck, P. D. Frederick, L. G. Shapiro, and J. G. Elmore. Region of interest identification and diagnostic agreement in breast pathology. *Modern Pathology*, 29(9):1004–1011, Sep 2016.

[77] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In S. Naik, editor, *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 284–287, May 2008.

[78] H. D. Nelson, M. Pappas, A. Cantor, J. Griffin, M. Daeges, and L. Humphrey. Harms of Breast Cancer Screening: Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation. *Annals of Internal Medicine*, 164(4):256–267, 2016.

[79] T. Newton, P. Slade, N. Butler, and P. Murphy. Personality and performance on a simple visual search task. *Personality and Individual Differences*, 13(3):381–382, Mar 1992.

[80] C. F. Nodine, H. Liu, W. T. Miller, and H. L. Kundel. Observer performance in the localization of tubes and catheters on digital chest images: the role of expertise and image enhancement. *Academic Radiology*, 3(10):834–841, 1996.

[81] H. Noh, S. Hong, and B. Han. Learning Deconvolution Network for Semantic Segmentation. In *2015 IEEE International Conference on Computer Vision*, pages 1520–1528, Dec 2015.

[82] K. C. Oeffinger, E. T. H. Fontham, R. Etzioni, A. Herzig, J. S. Michaelson, Y.-C. T. Shih, L. C. Walter, T. R. Church, C. R. Flowers, S. J. LaMonte, A. M. D. Wolf, C. DeSantis, J. Lortet-Tieulent, K. Andrews, D. Manassaram-Baptiste, D. Saslow, R. A. Smith, O. W. Brawley, and R. Wender. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. *JAMA*, 314(15):1599–1614, 2015.

[83] T. Onega, D. L. Weaver, B. Geller, N. Oster, A. N. A. Tosteson, P. A. Carney, H. Nelson, K. H. Allison, F. P. OMalley, S. J. Schnitt, and J. G. Elmore. Digitized Whole Slides for Breast Pathology Interpretation: Current Practices and Perceptions. *Journal of Digital Imaging*, 27(5):642–648, 2014.

[84] N. V. Oster, P. A. Carney, K. H. Allison, D. L. Weaver, L. M. Reisch, G. Longton, T. Onega, M. Pepe, B. M. Geller, H. D. Nelson, T. R. Ross, A. N. A. Tosteson, and J. G. Elmore. Development of a diagnostic test set to assess agreement in breast pathology: practical application of the Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *BMC Women's Health*, 13(1):3, Dec 2013.

[85] L. Pantanowitz, A. J. Evans, J. D. Pfeifer, L. C. Collins, P. N. Valenstein, K. J. Kaplan, D. C. Wilbur, and T. J. Colgan. Review of the current state of whole slide imaging in pathology. *Journal of Pathology Informatics*, 2(1):36, 2011.

[86] T. K. Paraïso, M. Kalaee, L. Zang, H. Pfeifer, F. Marquardt, and O. Painter. Position-Squared Coupling in a Tunable Photonic Crystal Optomechanical Cavity. *Physical Review X*, 5(4): 041024, Nov 2015.

[87] V. Raghunath, M. Braxton, S. Gagnon, T. T. Brunyé, K. H. Allison, L. Reisch, D. L. Weaver, J. G. Elmore, and L. G. Shapiro. Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. *Journal of Pathology Informatics*, 3(1):43, 2012.

[88] R. Randell, R. A. Ruddle, C. Mello-Thoms, R. G. Thomas, P. Quirke, and D. Treanor. Virtual reality microscope versus conventional microscope on time to diagnosis: an experimental study. *Histopathology*, In Press(2):1–18, 2007.

[89] X. Ren and J. Malik. Learning a classification model for segmentation. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 10–17, 2003.

[90] W. Road, N. Cary, and P. Charles. College of American Pathologists. `http://www.cap.org/apps/portlets/contentViewer/show.do?contentReference=cap_today%2F0112%2F0112a_regulators.html`, 2011.

[91] D. Romo, E. Romero, and F. González. Learning regions of interest from low level maps in virtual microscopy. *Diagnostic Pathology*, 6(Suppl 1):S22, 2011.

[92] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[93] J. Rosai. Borderline Epithelial Lesions of the Breast. *The American Journal of Surgical Pathology*, 15(3):209–221, Mar 1991.

[94] 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. *Computerized Medical Imaging and Graphics*, 35(7-8):603–615, Oct 2011.

[95] L. Roux, D. Racoceanu, N. Loménie, M. Kulikova, H. Irshad, J. Klossa, F. Capron, C. Genestie, G. Le Naour, and M. N. Gurcan. Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics*, 4(1):8, 2013.

[96] A. Saco, J. A. Bombi, A. Garcia, J. Ramírez, and J. Ordi. Current Status of Whole-Slide Imaging in Education. *Pathobiology*, 83(2-3):79–88, Apr 2016.

[97] 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. *Journal of Computational Science*, 3(5):269–279, Sep 2012.

[98] S. J. Schnitt, J. L. Connolly, F. A. Tavassoli, R. E. Fechner, R. L. Kempson, R. Gelman, and D. L. Page. Interobserver Reproducibility in the Diagnosis of Ductal Proliferative Breast Lesions Using Standardized Criteria. *The American Journal of Surgical Pathology*, 16(12): 1133–1143, Dec 1992.

[99] O. Sertel, J. Kong, H. Shimada, U. Catalyurek, J. H. Saltz, and M. N. Gurcan. Computer-aided prognosis of neuroblastoma: classification of stromal development on whole-slide images. *Pattern Recognition*, 6915(6):69150P, Mar 2008.

[100] O. Sertel, J. Kong, U. V. Catalyurek, G. Lozanski, J. H. Saltz, and M. N. Gurcan. Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading. *Journal of Signal Processing Systems*, 55(1):169–183, Apr 2009.

[101] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, Feb 2016.

[102] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, A. Böhm, O. Ronneberger, B. B. Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. Snead, and N. M. Rajpoot. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35: 489–502, Fan 2017.

[103] J. Sivic and A. Zisserman. Efficient Visual Search of Videos Cast as Text Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):591–606, Apr 2009.

[104] B. L. Sprague, R. E. Gangnon, V. Burt, A. Trentham-Dietz, J. M. Hampton, R. D. Wellman, K. Kerlikowske, and D. L. Miglioretti. Prevalence of Mammographically Dense Breasts in the United States. *JNCI Journal of the National Cancer Institute*, 106(10):dju255–dju255, Sep 2014.

[105] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 1–9, 2015.

[106] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi. Multifeature Prostate Cancer Diagnosis and Gleason Grading of Histological Images. *IEEE Transactions on Medical Imaging*, 26(10):1366–1378, Oct 2007.

[107] K. A. Thomas, M. J. Sottile, and C. M. Salafia. Unsupervised Segmentation for Inflammation Detection in Histopathology Images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6134 LNCS, pages 541–549. 2010.

[108] J. A. Tice, S. R. Cummings, R. Smith-Bindman, L. Ichikawa, W. E. Barlow, and K. Kerlikowske. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Annals of Internal Medicine*, 148(5):337–47, Mar 2008.

[109] G. Tourassi, S. Voisin, V. Paquit, and E. Krupinski. Investigating the link between radiologists' gaze, diagnostic decision, and image content. *Journal of the American Medical Informatics Association*, 20(6):1067–1075, Nov 2013.

[110] D. Treanor, C. H. Lim, D. Magee, A. Bulpitt, and P. Quirke. Tracking with virtual slides: a tool to study diagnostic error in histopathology. *Histopathology*, 55(1):37–45, 2009.

[111] N. Velez, D. Jukic, and J. Ho. Evaluation of 2 whole-slide imaging applications in dermatopathology. *Human Pathology*, 39(9):1341–1349, Sep 2008.

[112] M. Veta, P. J. van Diest, S. M. Willems, H. Wang, A. Madabhushi, A. Cruz-Roa, F. Gonzalez, A. B. Larsen, J. S. Vestergaard, A. B. Dahl, D. C. Cirean, J. Schmidhuber, A. Giusti, L. M. Gambardella, F. B. Tek, T. Walter, C.-W. Wang, S. Kondo, B. J. Matuszewski, F. Precioso, V. Snell, J. Kittler, T. E. de Campos, A. M. Khan, N. M. Rajpoot, E. Arkoumani, M. M. Lacle, M. A. Viergever, and J. P. Pluim. Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Medical Image Analysis*, 20(1):237–248, Feb 2015.

[113] J. Vink, M. van Leeuwen, C. van Deurzen, and G. de Haan. Efficient nucleus detector in histopathology images. *Journal of Microscopy*, 249(2):124–135, Feb 2013.

[114] T. Wan, X. Liu, J. Chen, and Z. Qin. Wavelet-based statistical features for distinguishing

mitotic and non-mitotic cells in breast cancer histopathology. In *2014 IEEE International Conference on Image Processing*, pages 2290–2294, Oct 2014.

[115] L. Wang and D.-C. He. Texture classification using texture spectrum. *Pattern Recognition*, 23(8):905–910, Fan 1990.

[116] G. Wen, A. Aizenman, T. Drew, J. M. Wolfe, T. M. Haygood, and M. K. Markey. Computational assessment of visual search strategies in volumetric medical images. *Journal of Medical Imaging*, 3(1):015501, 2016.

[117] M. Wilson. From processing efficiency to attentional control: a mechanistic account of the anxiety performance relationship. *International Review of Sport and Exercise Psychology*, 1(2):184–201, 2008.

[118] S. Wu, S. Zhong, and Y. Liu. Deep residual learning for image steganalysis. *Multimedia Tools and Applications*, pages 1–17, 2017.

[119] B. Xu, N. Wang, T. Chen, and M. Li. Empirical Evaluation of Rectified Activations in Convolutional Network. *ICML Deep Learning Workshop*, pages 1–5, May 2015.

[120] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang. Object Contour Detection with a Fully Convolutional Encoder-Decoder Network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–202, 2016.

[121] F. Yin, G. Han, M. M. Bui, J. Gibbs, I. Martin, L. Sundharkrishnan, L. King, C. Jabcuga, L. N. Stuart, and L. A. Hassell. Educational Value of Digital Whole Slides Accompanying Published Online Pathology Journal artciles: A Multi-Institutional Study. *Archives of Pathology & Laboratory Medicine*, 140(7):694–697, 2016.

[122] Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Graf, S. Chin, R. F. Schwarz, C. Curtis, M. J. Dunning, H. Bardwell, N. Johnson, S. Doyle, G. Turashvili, E. Provenzano, S. Aparicio, C. Caldas, and F. Markowetz. Quantitative Image Analysis of Cellular Heterogeneity in

Breast Tumors Complements Genomic Profiling. *Science Translational Medicine*, 4(157): 157ra143–157ra143, Oct 2012.

Appendix A

DIGIPATH STUDY FORMS AND DOCUMENTS

A.1 digiPATH Participant Baseline Survey

SURVEY OF PATHOLOGISTS

Instructions: **This survey takes < 10 minutes to complete. It asks about your background and what we think are extremely important general questions related to research and clinical care in breast pathology.**

GENERAL PROFESSIONAL INFORMATION

1. What is your year of birth? ___ ___ ___ ___ Year

2. What is your gender?
- ☐ Male
- ☐ Female

3. Are you affiliated with an academic medical center
- ☐ Yes, adjunct/affiliated clinical faculty
- ☐ Yes, primary appointment
- ☐ No

4. Have you received fellowship training in surgical or breast pathology? (check all that apply)
- ☐ Yes, surgical
- ☐ Yes, breast pathology
- ☐ No

5. The following questions are about your experience interpreting breast pathology cases.

a. How many years have you been interpreting breast pathology cases (not including residency/fellowship training)?
- ☐ < 1 year
- ☐ 1-2 years
- ☐ 3-4 years
- ☐ 5-9 years
- ☐ 10-19 years
- ☐ ≥20 years

b. What percentage of your caseload includes interpreting breast specimens?
- ☐ <10%
- ☐ 10-24%
- ☐ 25-49%
- ☐ 50-74%
- ☐ ≥75%

c. Estimate the number of breast cases you interpret during an average week.
- ☐ <5 breast cases per week
- ☐ 5-9 breast cases per week
- ☐ 10-19 breast cases per week
- ☐ 20-29 breast cases per week
- ☐ 30-39 breast cases per week
- ☐ 40-49 breast cases per week
- ☐ ≥50 breast cases per week

d. Do your colleagues <u>consider you an expert</u> in breast pathology?
☐ Yes
☐ No

6. In general, **<u>how challenging</u>** do you find breast cases to interpret?
1☐ 2☐ 3☐ 4☐ 5☐ 6☐
Very easy Very challenging

7. What are your thoughts on interpreting breast pathology?

	Strongly disagree 1	**Disagree 2**	**Slightly disagree 3**	**Slightly agree 4**	**Agree 5**	**Strongly agree 6**
A. Interpreting breast pathology is enjoyable	□	□	□	□	□	□
B. Interpreting breast pathology makes me more nervous than other types of pathology.	□	□	□	□	□	□

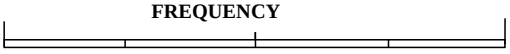
8. In general, **<u>how confident</u>** are you in your assessments of breast cases?

1☐ 2☐ 3☐ 4☐ 5☐ 6☐
Very confident Not at all confident

SECOND OPINION BY ANOTHER PATHOLOGST ON BREAST SPECIMENS (e.g. consultation, second read, second review)

9. Please consider the following hypothetical scenario…. You are reviewing a breast needle core biopsy from a 45 year old woman with no history of breast disease. There is an intra-ductal process that you consider to be borderline between atypical ductal hyperplasia (ADH) and ductal carcinoma in situ (DCIS), but you favor classifying as ADH.

a. In situations like this, what percentage of cases would you get a second opinion?

FREQUENCY
0%  100% ____%

(hover mouse cursor over bar to see percentage, or type a number in the box)

b. If you were to obtain a second opinion, would your second reviewer usually be blinded to your opinion on the case?
☐ Yes, they would be blinded
☐ No

c. If you obtain a second opinion and they favor DCIS, how often would you use the following methods to resolve the disagreement?

i. <u>Try to come to consensus</u> by discussing the case with the second reviewer

0% 100% ____%

☐ Not used

ii. Diagnose according to the <u>most experienced pathologist's opinion</u>

0% 100% ____%

☐ Not used

iii. <u>Get a third "tie-breaker"</u> opinion or present at a consensus conference

0% 100% ____%

☐ Not used

iv. Diagnose as <u>borderline</u> or <u>suspicious</u> (i.e. "ADH bordering on DCIS" or "ADH suspicious for DCIS")

0% 100% ____%

☐ Not used

v. Diagnose as DCIS to go with the <u>more severe</u> diagnosis

0% 100% ____%

☐ Not used

vi. Diagnose as ADH to go with the <u>less severe</u> diagnosis

0% 100% ____%

☐ Not used

vii. Other: _____

0% 100% ____%

☐ Not used

10. Some facilities have <u>policies requiring</u> a second opinion which may differ from our <u>actual practices</u> or what we think is <u>ideal for patient care</u>. Please describe your experience and thoughts on second opinions:

INITIAL DIAGNOSIS	**POLICY REQUIRED** (% of cases for which <u>my practice requires me</u> to obtain a second opinion)	**ACTUAL PRACTICE** (% of cases for which I <u>usually obtain</u> a second opinion)	**IDEAL PRACTICE FOR PATIENT CARE** (% of cases which I think <u>should ideally receive</u> a second opinion)
Invasive	**0%** ---------- **100%**	**0%** ---------- **100%**	**0%** ---------- **100%**

DCIS	**0% ---------- 100%**	**0% ---------- 100%**	**0% ---------- 100%**
ADH	**0% ---------- 100%**	**0% ---------- 100%**	**0% ---------- 100%**
Negative (non-atypical)	**0% ---------- 100%**	**0% ---------- 100%**	**0% ---------- 100%**

11. What are your thoughts on asking another pathologist for a second opinion on cases?

	DISAGREE			**AGREE**		
	Strongly disagree	**Disagree**	**Slightly disagree**	**Slightly agree**	**Agree**	**Strongly agree**
	1	**2**	**3**	**4**	**5**	**6**
A. Improves my diagnostic accuracy	□	□	□	□	□	□
B. Takes too much time	□	□	□	□	□	□
C. Protects me from malpractice suits	□	□	□	□	□	□
D. I wish it was more available	□	□	□	□	□	□
E. I'm often hesitant to request as it may make me look less adequate as a diagnostician	□	□	□	□	□	□

DIGITIZED WHOLE SLIDE IMAGING

(Virtual microscopy is a digital process by which an electronic scanner converts histological slides into high-resolution digitized pictures known as digitized whole slide images. The term "digitized whole slides" does not refer to jpeg-style images or PowerPoint images.)

12. In what ways do you use digitized whole slides in your professional work? (check all that apply)

- ☐ Primary pathology diagnosis
- ☐ Tumor board/clinical conference
- ☐ Consultative Diagnosis
- ☐ CME/Board exams/ Teaching in general
- ☐ Archival purposes
- ☐ Research
- ☐ Other: _____
- ☐ Not at all (skip to Question 14)

13.

13a. Do you interpret digitized whole H & E slide images of breast tissue for rendering a primary diagnosis?

- ☐ No
- ☐ Yes

POP UP if YES:
I. I render a primary diagnosis in _____% of my H & E breast cases using digital whole slide imaging.
II. I render a second opinion on _____% of my second review/consultation cases using digital whole slide imaging.

III. How long have you been using digital whole slide imaging for H & E interpretation of breast cases?
☐ *≤ 6 MONTHS* ☐ *> 6 MONTHS*

13b. Do you interpret digitized whole slide images on IHC stained breast tissue slides for rendering a primary diagnosis?
☐ No
☐ Yes

POP UP if YES: I. I interpret digitized whole IHC slides in breast cases for the following (check all that apply)
☐ *Prognostic/predictive breast cancer markers (e.g., ER, HER2, other)*
☐ *Diagnostic questions (e.g., Invasive cancer vs. DCIS, E-cadherin, other)*

14. What are your thoughts on H & E digitized whole slide imaging being used for primary diagnostic purposes? (We refer to digital whole slide images as digital slides)

	Strongly disagree 1	**Disagree 2**	**Slightly disagree 3**	**Slightly agree 4**	**Agree 5**	**Strongly agree 6**
A. Accurate diagnoses can be rendered using digital slides	□	□	□	□	□	□
B. Digital slides are useful for obtaining a second opinion	□	□	□	□	□	□
C. Digital slides increase pathologist exposure to medical malpractice suits	□	□	□	□	□	□
D. It is too difficult to learn how to use digital slides	□	□	□	□	□	□
E. Overall I think the benefits of digital whole slide imaging outweigh the concerns	□	□	□	□	□	□
F. Digital slides are too slow for routine use when interpreting a case	□	□	□	□	□	□
G. I would like to adopt digital whole slide imaging or increase use of it in my personal practice	□	□	□	□	□	□

MEDICAL MALPRACTICE

15. Have you ever been named in a medical malpractice suit (including any suit filed *and* either dropped, settled out of court or gone to trial)? (*check all that apply)*
☐ No, never been sued
☐ Yes, suit(s) related to breast pathology cases
☐ Yes, suit(s) related to other pathology or other medical cases

16. Have medical malpractice concerns affected **your peer's** practice with breast cases in the following ways?

	Strongly disagree	**Disagree**	**Slightly disagree**	**Slightly agree**	**Agree**	**Strongly agree**
A. My peers order additional immunohistochemistry tests	☐	☐	☐	☐	☐	☐
B. My peers recommend						

additional surgical sampling	☐	☐	☐	☐	☐	☐
C. My peers request additional reviews (second opinion)	☐	☐	☐	☐	☐	☐
D. When a case is borderline between DCIS and ADH, my peers generally choose the more severe diagnosis of DCIS	☐	☐	☐	☐	☐	☐

17. Have medical malpractice concerns affected **your own practice** with breast cases in the following ways?

	Strongly disagree	**Disagree**	**Slightly disagree**	**Slightly agree**	**Agree**	**Strongly agree**
A. I order additional IHC tests	☐	☐	☐	☐	☐	☐
B. I recommend additional surgical sampling	☐	☐	☐	☐	☐	☐
C. I request additional reviews (second opinion)	☐	☐	☐	☐	☐	☐
D. When a case is borderline between DCIS and ADH, I generally choose the more severe diagnosis of DCIS	☐	☐	☐	☐	☐	☐

CONTACT INFORMATION

We will contact you in the next few weeks to schedule a convenient time for your review of the sample breast cases. Below is the contact information we have for you. Please edit and/or complete this information as needed.

Daytime phone	(Auto populate)
Email	(Auto populate)
Evening phone	(Leave blank)
Cell phone	(Leave blank)
address	(Auto populate)
City	(Auto populate)
Zip code	(Auto populate)

☐ Click this box to confirm the above information is correct

The best way to reach me is (check all that apply)

☐ Email

☐ Daytime phone

☐ Evening phone

☐ Cell phone

Thank you for participating in this exciting study. Feel free to share any additional comments:

_____

Click here to submit your survey.

A.2 digiPATH Histology Form

Draft Date: October 29, 2010

Primary Key: _____
Pathologist Name: _____

Clinical History
Patient Age _____
Specimen Type: ☐ Core needle biopsy ☐ Excisional biopsy

I. Histologic Assessment: Diagnoses – Check all that apply. Choose the best fit among the options.

Non-Proliferative changes
- ☐ Non-proliferative changes only

Proliferative lesion without atypia:
- ☐ Fibroadenoma
- ☐ Intraductal papilloma without atypia
- ☐ Usual ductal hyperplasia
- ☐ Columnar cell hyperplasia /Columnar cell change
- ☐ Sclerosing adenosis
- ☐ Radial scar/complex sclerosing lesion

Atypical lesion:
- ☐ Flat epithelial atypia
- ☐ Atypical ductal hyperplasia
- ☐ Intraductal papilloma with atypia
- ☐ Atypical lobular hyperplasia

Carcinoma in situ:
- ☐ Ductal carcinoma in situ:

  Nuclear grade:
  a. ☐ Low
  b. ☐ Intermediate
  c. ☐ High

  Necrosis**:**
  a. ☐ Absent
  b. ☐ Present, focal (small foci/single cell necrosis)
  c. ☐ Present, central (expansive "comedo" necrosis)
- ☐ Lobular carcinoma in situ

(For mixed ductal & lobular features, check both DCIS & LCIS boxes and nuclear grade + necrosis)

Invasive carcinoma :
- ☐ Invasive carcinoma (ductal, lobular or other special type):
  - a. Tubule formation score: 1☐ 2☐ 3☐
  - b. Nuclear grade score 1☐ 2☐ 3☐
  - c. Mitotic activity score: 1☐ 2☐ 3☐
  - Overall Nottingham grade: ☐ Low(total score 3-5) ☐ Intermediate(6, 7) ☐ High(8, 9)

Additional comments: _____
_____
_____

II. If you considered this case borderline between two diagnoses, which diagnoses were you considering? Please check only two options: (Otherwise skip to Section III.)

Non-Proliferative changes
- ☐ Non-proliferative changes only

Proliferative lesion without atypia:
- ☐ Fibroadenoma
- ☐ Intraductal papilloma without atypia
- ☐ Usual ductal hyperplasia
- ☐ Columnar cell hyperplasia /Columnar cell change
- ☐ Sclerosing adenosis
- ☐ Radial scar/complex sclerosing lesion

Atypical lesion:
- ☐ Flat epithelial atypia
- ☐ Atypical ductal hyperplasia
- ☐ Intraductal papilloma with atypia
- ☐ Atypical lobular hyperplasia

Carcinoma in situ:

☐ Ductal carcinoma in situ:

☐ Lobular carcinoma in situ

Invasive carcinoma:

☐ Invasive carcinoma

What particular features made you favor the final diagnostic category you chose for the lesion?

_____

III. Additional questions regarding this case:

Please rate on the following scale your opinion of the **level of diagnostic difficulty** of this case:

1☐ 2☐ 3☐ 4☐ 5☐ 6☐

Very easy Very challenging

Please rate on the following scale your **confidence in your assessment**:

1☐ 2☐ 3☐ 4☐ 5☐ 6☐

Very confident Not at all confident

Would you ask for a **second pathologist's opinion** of this case before finalizing the report? (Assume a pathologist is available)

1. ☐ No
2. ☐ Yes, informal (not documented in report)
3. ☐ Yes, formal (documented in report)

IV. For 3-member expert panel only

Should this case be included in a test set?

1. ☐ No 2. ☐ Yes

Comments for expert panel to discuss:

Appendix B

CNNS FOR BREAST BIOPSY WSI SEGMENTATION

B.1 Definitions

B.1.1 Sum Fusion

A sum fusion function f^{sum} (eq B.1) computes the element-wise sum of two vector inputs $\mathbf{x}_a$ and $\mathbf{x}_b$ to produce output $\mathbf{y}_{sum}$.

$$f^{sum}(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{y}_{sum} = \mathbf{x}_a + \mathbf{x}_b \tag{B.1}$$

B.1.2 Convolutional fusion

A convolutional fusion function f^{conv} (eq B.2) first computes the element-wise sum of two vector inputs $\mathbf{x}_a$ and $\mathbf{x}_b$ using the sum fusion function f^{sum} and subsequently convolves the fused data with the bank of filters $\mathbf{W}$ and bias $\mathbf{b}$ to produce output $\mathbf{y}_{conv}$

$$f^{conv}(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{y}_{conv} = \mathbf{b} + f^{sum}(\mathbf{x}_a, \mathbf{x}_b) * \mathbf{W} \tag{B.2}$$

where $*$ is the convolution operator. For simplicity, we drop the bias term $\mathbf{b}$ in subsequent sections.

B.1.3 Encoder

The encoding network is a stack of encoding blocks that aggregates features at different spatial resolutions by performing the convolutional operation. A residual encoding function $E(\mathbf{x}_e)$ (eq B.3) convolves the input $\mathbf{x}_e$ with a bank of convolutional filters $\mathbf{W}_e$ and $\mathbf{W}_s$ to produce the output $\mathbf{y}_e$ as

$$E(\mathbf{x}_e) = \mathbf{y}_e = f^{sum}(\mathbf{x}_e * \mathbf{W}_e, \mathbf{W}_s * \mathbf{x}_e) \tag{B.3}$$

where $\mathbf{W}_s$ projects $\mathbf{x}_e$ into the vector space of $\mathbf{x}_e * \mathbf{W}_e$. Note that for non-residual encoding blocks, the encoding function is defined as $E(\mathbf{x}_e) = \mathbf{x}_e * \mathbf{W}_e$ with $\mathbf{W}_s = \mathbf{0}$.

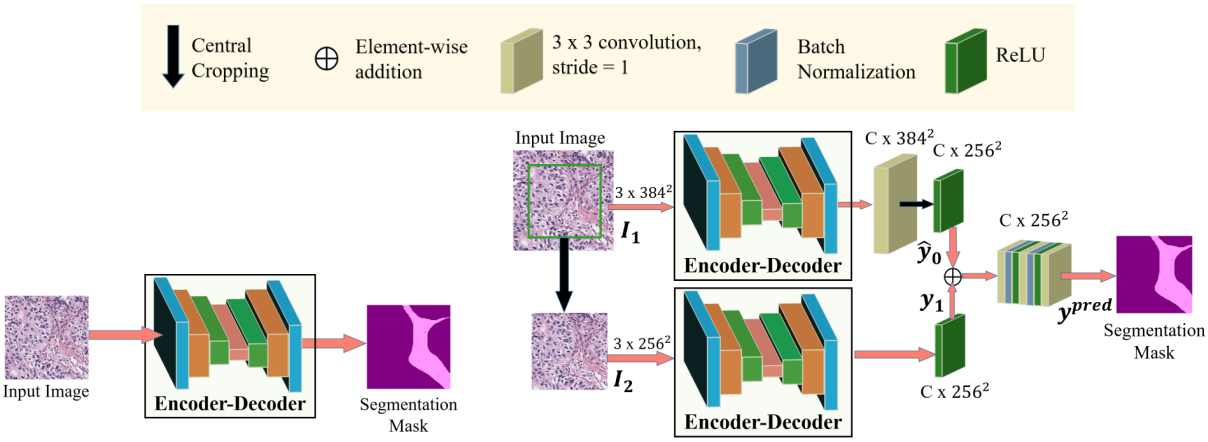
B.1.4 Decoder

The decoding network is a stack of decoding blocks that decodes the encoded feature map into the segmentation mask. The decoding function with skip-connection $D(\mathbf{x}_d, \mathbf{y}_e)$ (eq B.4) deconvolves the input $\mathbf{x}_d$ with a bank of deconvolutional filters $\mathbf{W}_d$ that learn the non-linear up-sampling operation to produce the output $\mathbf{y}_d$ as

$$D(\mathbf{x}_d, \mathbf{y}_e) = \mathbf{y}_d = f^{sum}(\mathbf{x}_d * \mathbf{W}_d, \mathbf{y}_e) \tag{B.4}$$

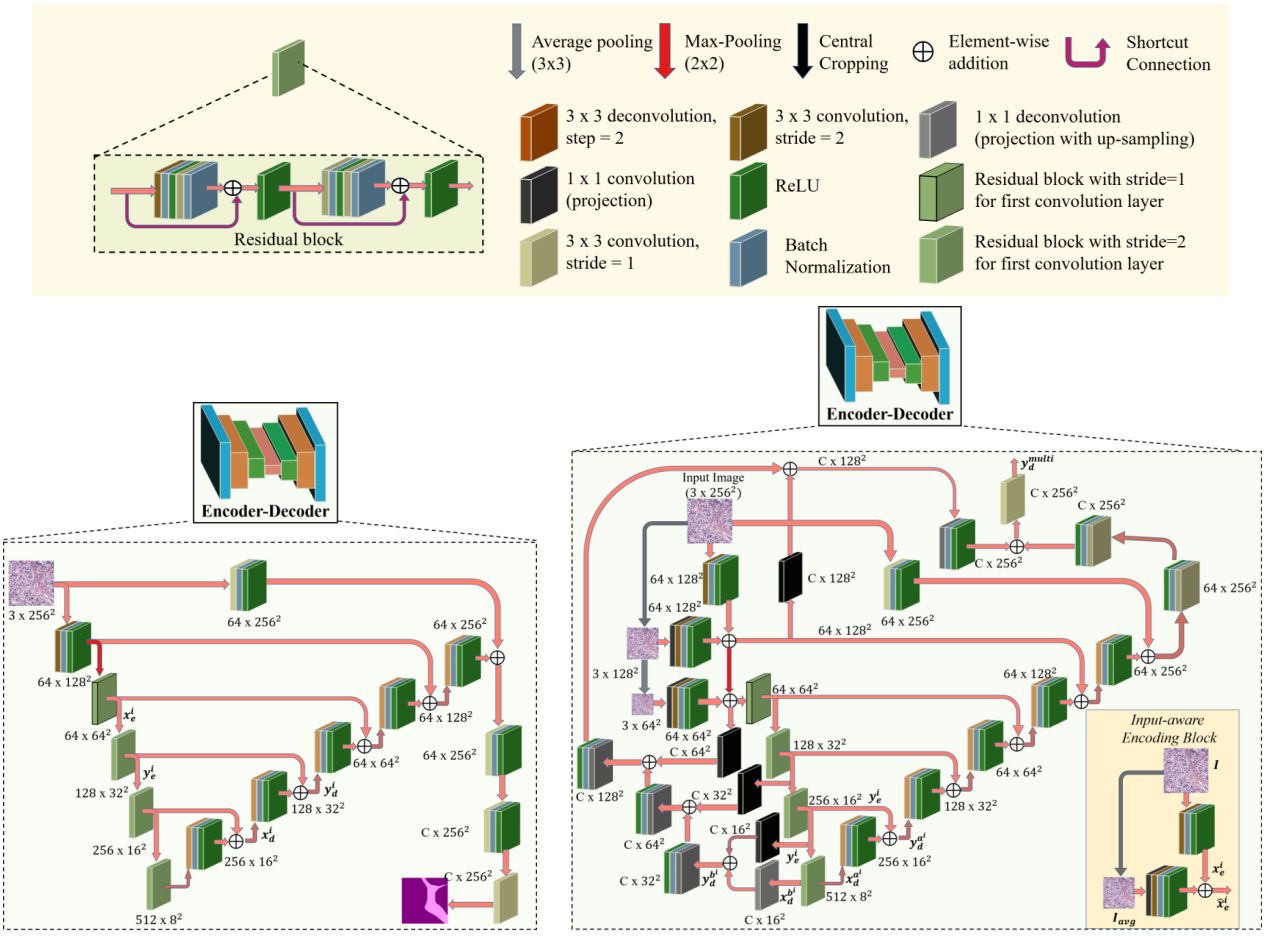
B.2 Plain-Network

Following previous work [36, 66, 81, 86, 92], we implemented a CNN-based segmentation method, *Plain-Network*, and found that such networks could not tackle these challenges. Plain-Network (Figure B.1a) is an encoder-decoder architecture. It uses the residual blocks and shares information between encoding and its corresponding decoding block using residual connections (Figure B.2a).



(a) Plain-Network (b) Multi-Resolution-Network

Figure B.1: Two network architectures: (a) Plain-Network, a simple encoder-decoder network, and (b) Multi-Resolution Network that combines multiple resolutions of the input image.



(a) Simple Encoder-Decoder

(b) Multi-Path Encoder-Decoder with Input-Aware Encoding Blocks

Figure B.2: Two encoder-decoder architectures: (a) A simple encoder-decoder architecture that is used in Plain-Network and MR-Plain-Network, (b) Multi-path ecoder-decoder architecture with input-aware encoding blocks that is used in MR-MP-Network.

B.3 Multi-Resolution Network

After analyzing the results of the Plain-Network we developed another network architecture, *Multi-Resolution Network* (Figure B.1b), that addresses some of the challenges our unique dataset presents. The main idea behind the Multi-Resolution Network is the aggregation of features from multiple resolutions of the input image. We used a larger patch as demonstrated in 5.9a together with the original patch to extract features at multiple resolutions. By aggregating the predictions of two encoder-decoder networks, we aimed to improve the segmentation especially at the borders of the patches.

The multi-resolution network consists of the composition of P instances of any encoding-decoding networks. Each instance learns the same number of parameters $\mathbf{W}$. The p^{th} instance is fed with the input patch $\mathbf{I}_p$ and generates the output $\mathbf{y}_p$ (eq B.5).

$$\mathbf{y}_p = D^{multi}(E^{aware}(\mathbf{I}_p)) \tag{B.5}$$

A cropping function $Cr(\mathbf{y}_p)$ (eq B.6) takes the output of the p^{th} instance and centrally crops it to produce the output $\hat{\mathbf{y}}_p$, which has the same dimensions as $\mathbf{y}_P$.

$$Cr(\mathbf{y}_p) = \hat{\mathbf{y}}_p \tag{B.6}$$

After cropping, a multi-resolution fusion function $Mr(\hat{\mathbf{y}}_p, \mathbf{y}_P)$ (eq B.7) is applied to fuse the output of these P network instances to produce the output $\mathbf{y}^{pred}$, which is aware of the context (or tissue information) surrounding the patch.

$$Mr(\hat{\mathbf{y}}_p, \mathbf{y}_P) = \mathbf{y}^{pred} = f^{conv}(\mathbf{y}_P, \sum_{p \neq P} \hat{\mathbf{y}}_p) \tag{B.7}$$

The network is then trained by minimizing the cross-entropy loss between the predicted value $\mathbf{y}^{pred}$ and the ground truth label $\mathbf{y}^{gth}$.

B.4 Multi-Path Encoder-Decoder with Input Aware Encoding Blocks

In addition to the multi-resolution network, we built a more powerful encoder-decoder architecture than that of Plain-Network, *Multi-Path Encoder-Decoder* (Figure B.2b), using two main ideas: efficient decoding of the contextual information at multiple levels and an input-aware information sharing in encoder. Following previous work [66, 120], we use a light weight decoder instead of a symmetric decoder [36]. We used deconvolution for learning the non-linear up-sampling operation[66] to build our decoding block which consists of a 3×3 deconvolutional layer followed by a batch normalization layer [48] and an activation layer ReLU [119]. Apart from the up-sampling operation, the deconvolutional layer also reduces the feature map dimensions.

The following sections describe these modifications in detail.

B.4.1 Input-Aware Encoding Blocks

Spatial information, lost due to down-sampling operations (such as max-pooling), is hard to be recovered by the decoder. Skip-connections help to improve the information flow between the encoder and decoder, thereby improving overall network accuracy. To further improve the information flow between encoder and decoder, we introduce an *input-aware encoding block*. The proposed block finds important features with respect to the input image and the output of the previous encoding block before sharing it with the next encoding block. During the initial phase of training, the proposed method amplifies the features in the encoding block arbitrarily; however, as the network is optimized, it finds more relevant features, thus helping to improve the information flow between the encoder and decoder network.

Let $\mathbf{I}$ be the input tensor and $\mathbf{x}_e^i$ be the input of the i^{th} encoding block. We first down-sample the input tensor $\mathbf{I}$ using an average pooling operation, so that the input tensor has the same spatial dimensions as the input to the i^{th} encoding block $\mathbf{x}_e^i$. Let $\mathbf{I}_{avg}$ be the down-sampled input tensor. The input-aware encoding function E^{aware} takes the down-sampled input image $\mathbf{I}_{avg}$, projects it into the same dimensional vector space as $\mathbf{x}_e^i$ using a bank of filters $\mathbf{W}_{Pr}$, computes dense features using a bank of filters $\mathbf{W}_{De}$, and then combines dense features with the input of encoding block $\mathbf{x}_e$

to produce the output $\mathbf{y}_e$ as

$$\mathbf{y}_e^i = E^{aware}(\mathbf{x}_e^i, \mathbf{I}_{avg}) = E(\overbrace{f^{sum}(\mathbf{x}_e^i, (\mathbf{I}_{avg} * \mathbf{W}_{Pr}) * \mathbf{W}_{De})}^{\text{Input-aware fusion } (\hat{\mathbf{x}}_\mathbf{e}^\mathbf{i})}) \tag{B.8}$$

B.4.2 Multi-Path Decoding Network

We aim to efficiently combine low-level features of the encoding network with high-level features to generate a semantic segmentation mask for the given input patch $\mathbf{I}$. To do so, we must invert the loss of resolution from down-sampling. Using previous work [36, 81, 86, 92], we augment the encoder network with the bottom-up refinement approach. We introduce two decoding networks that decode the encoded input into a C-dimensional output, where C represents the number of classes in the dataset.

Let D^a and D^b denote the decoding functions corresponding to these decoding networks. The function D^a (eq B.9) deconvolves the input $\mathbf{x}_d^a$ with a bank of filters $\mathbf{W}_d^a$ to produce the output $\mathbf{y}_d^a$. The function D^a is the same as the decoder function (eq B.4) defined in Section B.2.

$$D^a(\mathbf{x}_d^a, \mathbf{y}_e) = \mathbf{y}_d^a = f^{sum}(\mathbf{x}_d^a * \mathbf{W}_d^a, \mathbf{y}_e) \tag{B.9}$$

The function D^b (eq B.10) projects the vectors $\mathbf{x}_d^b$ and $\mathbf{y}_e$ into C-dimensional vector space using a bank of 1×1 deconvolutional filters $\mathbf{W}_D$ and convolutional filters $\mathbf{W}_C$; it then fuses them using sum fusion to produce the output $\mathbf{y}_d^b$. The decoding functions D^a and D^b are repeated n-times until the spatial dimensions of feature maps $\mathbf{y}_d^a$ and $\mathbf{y}_d^b$ are the same as the input image.

$$D^b(\mathbf{x}_d^b, \mathbf{y}_e) = \mathbf{y}_d^b = f^{sum}(\mathbf{x}_d^b * \mathbf{W}_D, \mathbf{y}_e * \mathbf{W}_C) \tag{B.10}$$

The function D^{multi} (eq B.11) fuses the output of these two different networks using convolutional fusion to produce the output $\mathbf{y}_d^{multi}$. Note that $\mathbf{W}_D$ performs an up-sampling operation while simultaneously projecting the feature map to C-dimensional vector space.

$$D^{multi}(\mathbf{y}_d^{a^n}, \mathbf{y}_d^{b^n}) = \mathbf{y}_d^{multi} = f^{conv}(\mathbf{y}_d^{a^n}, \mathbf{y}_d^{b^n}) \tag{B.11}$$

B.4.3 Extension to Seven Label Classes

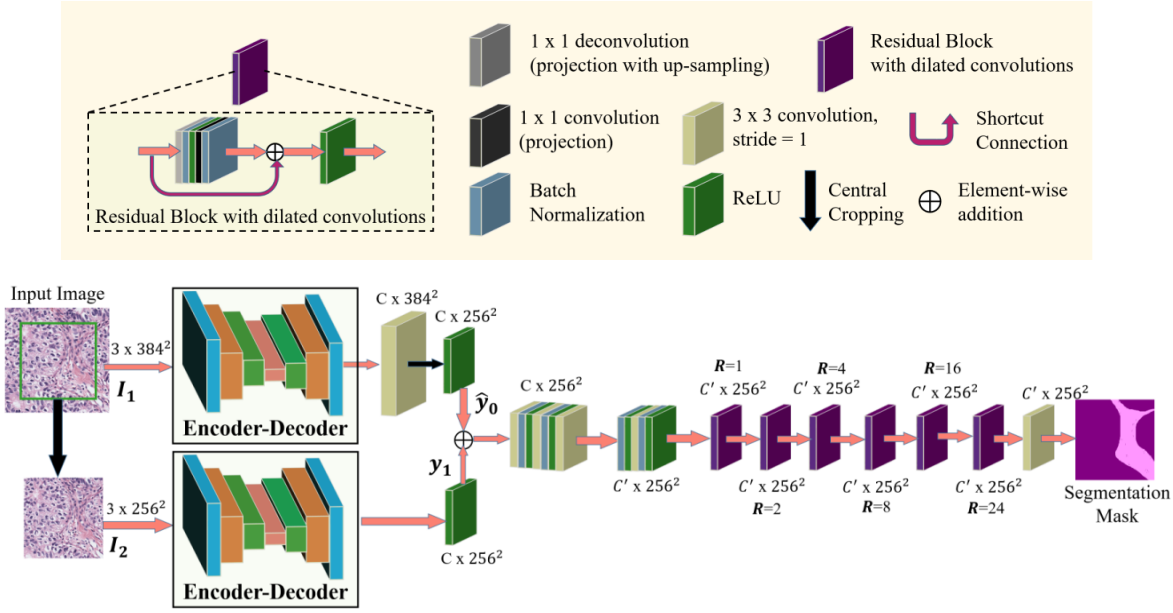


Figure B.3: The extension of MR-MP-Network to segment eight label classes and the context module developed using residual blocks with dilated convolutions.

Appendix C

ADDITIONAL ANALYSIS OF DIAGNOSTIC CLASSIFICATION

Table C.1: Precision and recall values of the mid-level and high-level features with different tissue histograms and segmentations for the diagnostic classification tasks.

Feature	*Segmentation*							
	precision				*recall*			
	SVM5		*SVM8*		*CNN5*		*CNN8*	
Invasive vs. Benign-Atypia-DCIS								
Superpixel label freq. & cooc. hist.	.67	.12	.87	.13	.65	.16	.57	.24
Superpixel label freq. & cooc. hist. (no bg)	.72	.09	.79	.15	.53	.22	.43	.70
Superpixel label freq. & cooc. hist. (no bg or str)	.89	.09	.71	.12	1.00	.11	.72	.17
Structure feature	.88	.11	.80	.17	.59	.18	.57	.49
Structure feature (no bg)	.86	.10	.77	.15	.53	.18	.59	.49
Structure feature (no bg or str)	.78	.15	.71	.16	1.00	.12	.72	.16
Atypia-DCIS vs. Benign								
Superpixel label freq. & cooc. hist.	.63	.79	.81	.79	.50	.80	.39	.88
Superpixel label freq. & cooc. hist. (no bg)	.54	.81	.70	.80	.46	.82	.24	.94
Superpixel label freq. & cooc. hist. (no bg or str)	.79	.79	.73	.83	.77	.78	.50	.94
Structure feature	.72	.80	.71	.85	.61	.82	.42	.93
Structure feature (no bg)	.80	.80	.72	.85	.60	.84	.42	.95
Structure feature (no bg or str)	.70	.81	.70	.86	.74	.79	.51	.93
DCIS vs. Atypia								
Superpixel label freq. & cooc. hist.	.48	.59	.69	.73	.48	.62	.98	.67
Superpixel label freq. & cooc. hist. (no bg)	.70	.60	.98	.62	.69	.60	.91	.75
Superpixel label freq. & cooc. hist. (no bg or str)	.43	.59	.87	.70	.87	.68	.81	.88
Structure feature	.84	.67	.86	.80	.61	.68	.82	.88
Structure feature (no bg)	.80	.63	.87	.76	.29	.78	.83	.89
Structure feature (no bg or str)	.71	.68	.82	.76	.91	.66	.81	.89

Table C.2: Sensitivity and specificity values of the mid-level and high-level features with different tissue histograms and segmentations for the diagnostic classification tasks.

Feature	*Segmentation*							
	sensitivity				*specificity*			
	SVM5		*SVM8*		*CNN5*		*CNN8*	
Invasive vs. Benign-Atypia-DCIS								
Superpixel label freq. & cooc. hist.	.12	.95	.13	.97	.16	.96	.24	.95
Superpixel label freq. & cooc. hist. (no bg)	.09	.93	.15	.97	.22	.95	.70	.95
Superpixel label freq. & cooc. hist. (no bg or str)	.09	.95	.12	.95	.11	1.00	.17	.96
Structure feature	.11	.97	.17	.97	.18	.95	.49	.96
Structure feature (no bg)	.10	.96	.15	.97	.18	.95	.49	.96
Structure feature (no bg or str)	.15	.97	.16	.96	.12	1.00	.16	.96
Atypia-DCIS vs. Benign								
Superpixel label freq. & cooc. hist.	.79	.34	.79	.41	.80	.31	.88	.33
Superpixel label freq. & cooc. hist. (no bg)	.81	.33	.80	.37	.82	.32	.94	.31
Superpixel label freq. & cooc. hist. (no bg or str)	.79	.41	.83	.42	.78	.37	.94	.39
Structure feature	.80	.38	.85	.44	.82	.36	.93	.36
Structure feature (no bg)	.80	.43	.85	.45	.84	.37	.95	.36
Structure feature (no bg or str)	.81	.38	.86	.44	.79	.37	.93	.39
DCIS vs. Atypia								
Superpixel label freq. & cooc. hist.	.59	.47	.73	.63	.62	.49	.67	.93
Superpixel label freq. & cooc. hist. (no bg)	.60	.51	.62	.92	.60	.52	.75	.85
Superpixel label freq. & cooc. hist. (no bg or str)	.59	.47	.70	.76	.68	.74	.88	.78
Structure feature	.67	.70	.80	.81	.68	.56	.88	.79
Structure feature (no bg)	.63	.61	.76	.80	.78	.50	.89	.80
Structure feature (no bg or str)	.68	.61	.76	.75	.66	.78	.89	.78

VITA

Ezgi Mercan recieved her Bachelor of Science in Computer Science from Bilkent University in Turkey, her Master of Science in Computer Science and Engineering from University of Washington in Seattle, WA. She is currently a PhD candidate in Computer Science and Engineering at the University of Washington.