HISTOPATHOLOGICAL IMAGE ANALYSIS WITH CONNECTIONS TO GENOMICS

BY

BENJAMIN CHIDESTER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Minh Do, Chair
Professor Jian Ma, Carnegie Mellon University
Professor Rohit Bhargava
Professor Stephen Boppart
Professor Zhi-Pei Liang

ABSTRACT

The fields of imaging and genomics in cancer research have been mostly studied independently, but recently available datasets have made investigation into the synergy of these two fields possible. This work demonstrates the efficacy of computational histopathological image analysis to extract meaningful quantitative nuclear and cellular features from hematoxylin and eosin stained images that have meaningful connections to genomic data. Additionally, with the advent of whole slide images, significantly more data representing the variation in nuclear characteristics and tumor heterogeneity is available, which can aid in developing new analytical tools, such as the proposed convolutional neural network for nuclear segmentation, which produces state-of-the-art segmentation results on challenging cases seen in normal pathology. This robust segmentation tool is essential for capturing reliable features for computational pathology. Additionally, whole slide images capture rich spatial information about tumors, which presents a challenge, but also an opportunity for the development of new image processing tools to capture this spatial information, which could be considered for future work. Other histopathological image modalities and relevant machine learning tools are also considered for elucidating cellular processes of cancer.

The fear of the LORD is the beginning of wisdom.
- Psalm 111:10

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my advisor, Prof. Minh Do. Though personal qualities of aptitude and perseverance are necessary in graduate school, just as necessary is the support, vision, and enthusiasm of an advisor, and I am thankful to have received these from Prof. Do throughout my academic career, along with his sincere care for my academic success and personal flourishing. As well, I am thankful for the crucial guidance and investment from Prof. Jian Ma, who made the integration of genomics in this research possible. I am also thankful for encouragement in this research from my committee: Prof. Stephen Boppart, Prof. Rohit Bhargava, and Prof. Zhi-Pei Liang.

What has made my graduate career truly enjoyable and meaningful has been the opportunity to have known many great friends through the ECE department and my research group: Trong Nguyen, Patrick Johnstone, Matthew Kole, Xinqi Qu, Ramanpreet Singh, Gregory Meyer, Luke Pfister, Jonathan Ligo, Honghai Yu, Teck Yian Lim, Chen Chen, Raymond Yeh, Vaishnavi Subramanian, Nguyen Mac, Dario Aranguiz, Siying Liu, Huy Bui, and many others.

There are several collaborators who made valuable contributions to this work whom I wish to acknowledge: Jack P. Hou for helpful discussions about cancer genomics; Chang Hu for help in the laborious task of collecting nuclei training samples; Vaishnavi Subramanian for helping to extend the proposed segmentation algorithm to lung cancer images; Andrew Bower for inviting me to work on the interesting problem of multimodal image analysis; several members of Prof. Bhargava's group for providing access to data; and Sadhya Sarwate, M.D. for helpful consultations about cancer pathology.

Surpassing all other relationships by which I have been blessed and which have encouraged me in my academics is my family. The love of my Mom and Dad, and of the rest of my family, and their support and encouragement

have given me the confidence to persevere and the security of knowing that I am loved regardless. I am also thankful for wonderful relationships through All Souls Presbyterian Church, the Graduate Christian Fellowship, and other ministries on campus that have deeply enriched my life at UIUC.

Lastly, I owe all thanks to my Lord and Savior Jesus Christ, for creating me with the skills and abilities for this work and for meriting on my behalf the peace that surpasses all understanding. It is for him and for his praise that I labor.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Cancer remains one of the top causes of death among diseases in the U.S. and the world. Joe Biden, former Vice President of the U.S., recently made an ambitious call for an earnest investment from federal government funding agencies to promote significant advances in cancer research, what he coined as the "cancer moonshot initiative," with the goal of curing cancer in the next few years.

Making such strides will require an interdisciplinary effort and an integrated understanding from the many fields of cancer research, primarily genomics and imaging. Imaging and genomics are two separate streams of inquiry into the cellular processes within a tumor and have been predominantly explored independently. In order for healthy cells to evolve into cancerous cells, form tumors, and ultimately to become malignant, they must acquire several characteristic abilities, including proliferative signaling, evasion of growth suppressors, activation of invasion and metastasis, replicative immortality, induction of angiogenesis, and resistance to cell death [1]. These problematic traits are caused by biological dysfunction at the genetic, molecular, and cellular levels and involve complex relationships between abnormal DNA, gene expression, protein synthesis, and even inter-cellular signaling. Different modalities of cancer data - from genomics, radiology, histopathological imaging - provide different quantitative and qualitative information about the state and development of these processes.

Recently, large-scale cancer genomic datasets have been collected along with histology hematoxylin and eosin (H&E) images of patients, such as The Cancer Genome Atlas (TCGA) [2], which also includes radiology images, and METABRIC [3], which has spurred new investigations and questions into the synergy of imaging-genomics research [4]. These histology slides are not biopsies, which capture only a small local area, but whole slide images (WSIs) of a slice of the entire tumor. These WSIs convey details about the

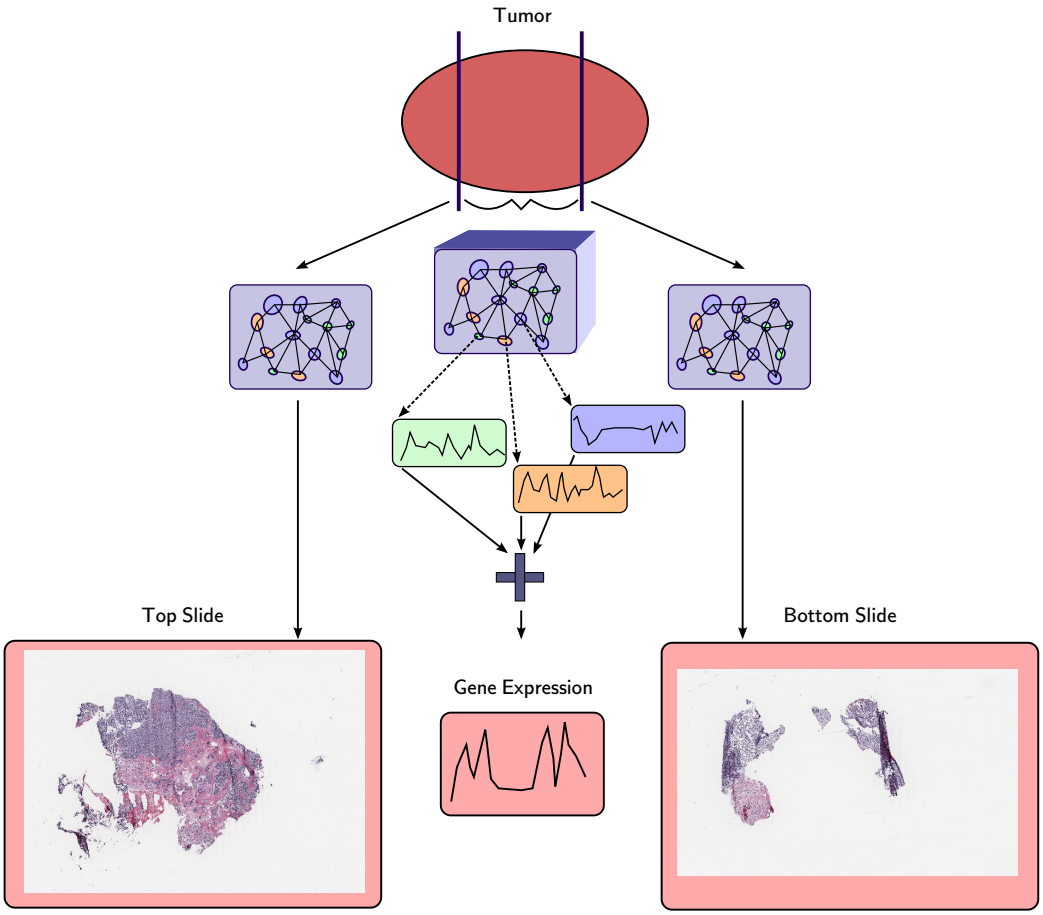


Figure 1.1: Diagram for acquisition of top and bottom H&E slides and gene expression data from a tumor.

tumor that cannot be observed from genomic data, such as its heterogeneity, the spatial relationships of particular cells, its composition of different cell types, and morphological features of cells and nuclei. These features can aid inference from genomic data, such as using lymphocyte counts to fix copy number variation measurements [5], or they can be used along with genomic data as additional indicators in computational prognostic models. Moreover, several of the slides that are imaged are extracted from the boundary of the tumor section that is used for genomic data extraction, meaning that the cells and their spatial layout in both the image slices and the section for genomics should closely match (see Fig. 1.1), though the validity of this assumption is highly dependent upon the type of cancer.

Several prominent imaging-genomic investigations have already produced significant findings. Recent understanding of the association of lymphocytic

infiltration with survival in breast cancer tumors [6] has spurred research into computational metrics to quantify lymphocyte, cancer cell, and stromal cell heterogeneity within a tumor and its association with survival [7, 8]. Incorporating the presence of a TP53 mutation into the analysis led to even stronger stratification of patients by survival [9]. Another investigation, considering only triple negative breast cancer (TNBC) patients, computed morphological features of H&E images, correlated them with discovered metagenes of gene expression, and showed the prognostic significance of these joint biomarkers [10].

Glioblastoma multiforme (GBM), also referred to as glioblastoma, has also received attention in imaging-genomic research. GBM is a high-grade astrocytoma, which is a brain tumor that originates in glial cells called astrocytes. Though rare among all cancers, it is the most common malignant brain tumor in adults and has poor prognosis due to the current lack of understanding among researchers. Genomic research has identified four molecular subtypes of GBM [11] (proneural, neural, classical, and mesenchymal), and recent work has attempted to classify these molecular types using computational image features of nuclei [12]. Subsequent work correlated discovered nuclear clusters with genomic data, and found significant associations between the clusters with alterations of several genes: EGFR amplification, CDKN2A deletion samples, and PTEN deletion samples [13]. Other work performed a similar procedure of clustering nuclear features in GBM and associating the clusters with gene expression and discovered six significantly associated regulatory hubs: IFNG, TGFB1, MAPK14, cytokines, PKC, and IL1B. A more recent work has investigated connections between quantitative metrics of necrosis and gene expression in GBM [14].

Though research has already begun toward this aim of joint imaging-genomic analysis, there are still many possible connections to be explored with the data available to the public research community and much more to be understood about the mapping from genotype to phenotype. Most analysis of TCGA has considered only modestly sized subsets of patients. Additionally, handling the sheer size of WSIs and quantifying the rich spatial information and cellular diversity that they capture remains a challenge [15]. A recent workshop organized by the National Cancer Institute on imaging-genomics echoed this challenge, identifying several important directions for future research, including the "development of more sophisticated imaging

methods to characterize this multi-cellular structure and how the microenvironment influences tumor behavior" and "image analysis methods to predict and detect the emergence of resistance, correlate with genomic heterogeneity, and identify homogeneous subtypes within a heterogeneous tumor" [16]. New insights will require domain-specific inference tools for both histopathological imaging and genomics and understanding of appropriate modeling of their connections.

1.1 Contributions

Motivated by the promise of imaging-genomics for furthering cancer research, the availability of large datasets of joint histopathological image and genomic data, and the need for computational inferential tools to meet the posed challenges, this thesis proposes several tools to aid in this research.

1.1.1 H&E Analysis Pipeline and Nuclei Segmentation

The primary contribution of this thesis is a robust pipeline for transforming H&E WSIs into a single feature vector composed of statistics of nuclear and cellular features. When cells become cancerous, they proliferate, forming irregular spatial patterns, and their nuclei become deformed. The proposed pipeline is able to capture these irregularities quantitatively by first segmenting the nuclei and cells in the image, providing their location and boundary, and then computing measures of their shape, size, color, and texture. Tools exist that attempt to extract such measurements, a popular example being CellProfiler [17], a software tool for analyzing microscopy image data, but these tools can produce unreliable measurements, especially when applied to a diversity of H&E images of varying quality.

The proposed pipeline uses CellProfiler for calculating measurements of segmented nuclei and cells, but replaces the standard methods of segmentation with a convolutional neural network (CNN) approach. Though the size of WSIs and the variation in quality of H&E images in large datasets, such as TCGA, presents a formidable computational challenge, this challenge also carries an advantage, in that much more of the variation of cells, nuclei, and their layout can be observed. A single WSI likely contains tens of thousands

of cells and nuclei, which can be used to capture their inherent variation in the training of novel data-driven approaches. Specifically, the task of nuclear segmentation, which is fundamental to H&E analysis, can be greatly advanced by new learning-based methods that leverage the diversity of nuclei in WSIs. With the added accuracy of nuclear and cellular masks produced by the CNN, this pipeline is able to extract reliable features, even with the challenging degree of diversity of staining and slicing present in TCGA H&E image data, in comparison to other pipelines often used in H&E analysis.

1.1.2 Genomic Integration

Using the proposed pipeline, this thesis demonstrates the utility of the computed features in stratifying cancer patients by outlook. An investigation into TCGA breast cancer data is presented, revealing image-based features that significantly separate patients into poor and improved prognosis groups using unsupervised clustering. The gene expression data of the patients is then tested for significant differential expression across the two image-based clusters, revealing 255 significant genes. Pathway analysis revealed that these genes were significantly associated with several pathways implicated in cancer.

Continued investigation using the proposed pipeline and corresponding image features will hopefully lead to the discovery of homogeneous subtypes within tumors, thereby aiding doctors in treatment assessment, similar to the clinically relevant discovery of four main subtypes in breast cancer from genomic data [18]. Potential outcomes could be the development of imaging surrogates for genetic mutations or expression, the discovery of new subtypes of cancer, and more accurate computational prognosis through the leveraging of joint genomic and imaging data. Furthermore, while work in computational pathology has revealed new insights into prognostic indicators, such as the importance of stromal cell features [19], the integration of genomic data with image features could lead to greater increases in prognostic accuracy and yet new insights.

1.1.3 Multimodal Histopathological Image Analysis

Although inspection of H&E histopathological images is still the gold standard for cancer diagnosis, other histopathological imaging modalities exist and are being developed that could be added to this investigation to provide further elucidation of phenomena in cancer.

When a tumor biopsy is taken, contiguous slices can be purposed for H&E and hyperspectral imaging modalities as well. Fourier transform infrared spectroscopy [20, 21] provides quantitative molecular vibration information of tissue without contaminating the tissue with stains. The response of the tissue at each frequency is a function of specific known molecular characteristics, providing more information than can be obtained by visible light imaging of H&E-stained tissue. Additionally, the spectrum of each pixel can be classified according to cell type, providing a pathologist with a visual map of the layout of the cells in the image, similar to that of H&E. Hyperspectral fluorescence lifetime imaging (FLIM) [22, 23] is another developing hyperspectral image modality, which measures not only the fluorescence intensity response of certain molecules, such as nicotinamide adenine dinucleotide (NADH), in tissue, but also the rate of decay of the response, which is informative of metabolic processes. A primary advantage of these modalities is that they are *label free*; that is, they require no staining or other contamination of the imaged tissue. Future work could investigate further connections between imaging and genomics by incorporating analysis of these hyperspectral modalities with H&E observations and genomic data.

Imaging devices are also being developed to image with several distinct modalities that are temporally and spatially registered. For example, optical coherence microscopy (OCM) provides structural information about the environment within a tumor, especially the boundaries of cell nuclei. FLIM can complement OCM by providing information about the metabolic activity within cells, which can be discriminative of different cellular processes in tumors, such as apoptosis and necrosis, the different methods of cell death. Additionally, images from these modalities can be captured over a time series to provide insight into the dynamic processes inside the microenvironment of tumors and within individual cells, and they can be employed *in vivo* for use in live studies and clinical applications. This thesis will demonstrate quantitatively the value of multimodal data over either modality by itself

and propose a formulation for the optimal detection rule for differentiating these processes. These imaging devices could be crucial in the future for investigating not only the impact of genomics on the morphology of cells and nuclei, but also on their movement and development over time as a tumor progresses.

CHAPTER 2

H&E ANALYSIS PIPELINE

In order to infer meaningful imaging-genomic connections, proper analysis of H&E histopathological images must be considered. H&E histopathological images are the gold standard for pathologists for diagnosis and prognosis of cancer, as they provide a clear, observable description of cells and nuclei and their spatial layout. Most cancer types have an established, methodical system for grading a tumor that pathologists follow that correlates with patient prognosis. These grading systems usually involve measurements of nuclear pleomorphism, such as shape, size, and granularity; counts of mitosis; and the spatial layout of epithelial cells, specifically, the presence or absence of clearly formed tubules.

Histopathological image analysis techniques have been actively researched over the past several decades [24, 25, 26]. Most work has focused on detecting and segmenting nuclei and describing their characteristics with quantitative metrics [27], driven by the intuition of conventional cancer grading. Some research in computational histopathological image analysis has even considered additional, unconventional metrics and has suggested their utility in cancer grading. A well-known recent finding from such research is the purported impact of features of stromal nuclei in breast cancer [19]. Computational pathology analysis also carries the benefit of avoiding inter-observer subjectivity and the promise of developing objective methods for diagnosis and prognosis. The availability of large datasets of histopathological images, such as TCGA, and the advent of WSIs have engendered a resurgent interest in computational pathology [28].

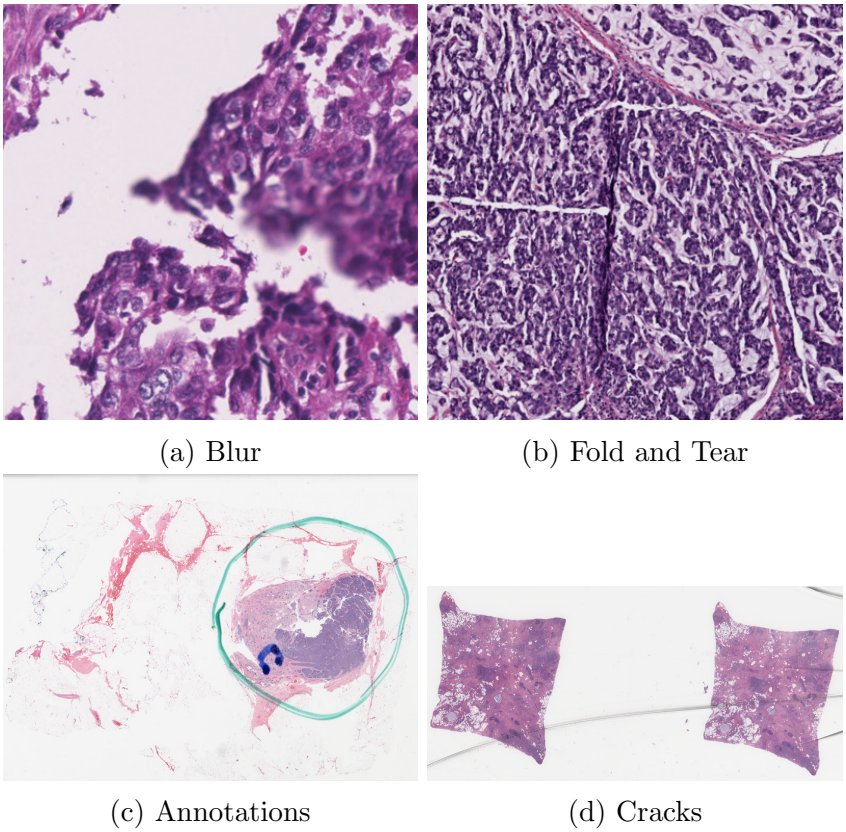
Challenges in developing robust computational metrics arise primarily from the acquisition process of H&E images. The tumor tissue specimen to be imaged is first fixed, usually with formaldehyde, to stabilize proteins and stop biological activity. The specimen is then processed so that it can be sectioned, which involves dehydrating the tissue, clearing the dehydrant, and

embedding the tissue, usually achieved with paraffin. Once embedded, the specimen can be sectioned into thin slices for imaging. A significant cause of variation in H&E images is slice thickness, which can vary roughly from 3 to 10 μm, causing significant variation in the intensity of subsequent staining. Additionally, if not properly processed, the sections can be difficult to slice, resulting in tearing artifacts. After the tissue is sliced, the process of embedding must be reversed for each slice so that it will receive the stain. Hematoxylin and eosin dyes are then applied either "regressively" or "progressively" to achieve a final stained sample for imaging. Hematoxylin stains nuclear regions blue or purple, while its counter-stain, eosin, stains cytoplasmic proteins pink. Over-staining with hematoxylin can cause a loss of detail of nuclei, while under-staining will lead to nuclei that are faint and not easily differentiable. If the paraffin was not properly removed prior to staining, the tissue will not take the stain, resulting in faint and blotchy images. Finally, the tissue section is mounted, during which other additional artifacts can be introduced, such as tissue folds or the presence of mounting mediums on the cover slip, resulting in image blurring. Several example artifacts from TCGA breast cancer images that were analyzed for this thesis are shown in Fig. 2.1. Image processing techniques must be employed that are able to mitigate the effects of such artifacts and remove uninformative variation.

The proposed pipeline for computing inference on H&E images consists of the following stages of processing: pre-processing, to unmix and normalize the stain; nuclei segmentation; cell segmentation, using discovered nuclei as seeds; cell and nuclei feature extraction; and feature summarization. Other possible stages that could be included are cell and nuclei classification, generally into stromal and epithelial cells and lymphocytes, and spatial feature extraction. Although these stages are not implemented currently in the proposed pipeline, they could be easily integrated, and future work will consider the merits of doing so.

2.1 Pre-Processing

Several pre-processing steps should be taken before analysis of H&E-stained slides, especially WSIs, to mitigate the effects of the aforementioned artifacts and uninformative variation. These steps usually include stain normalization



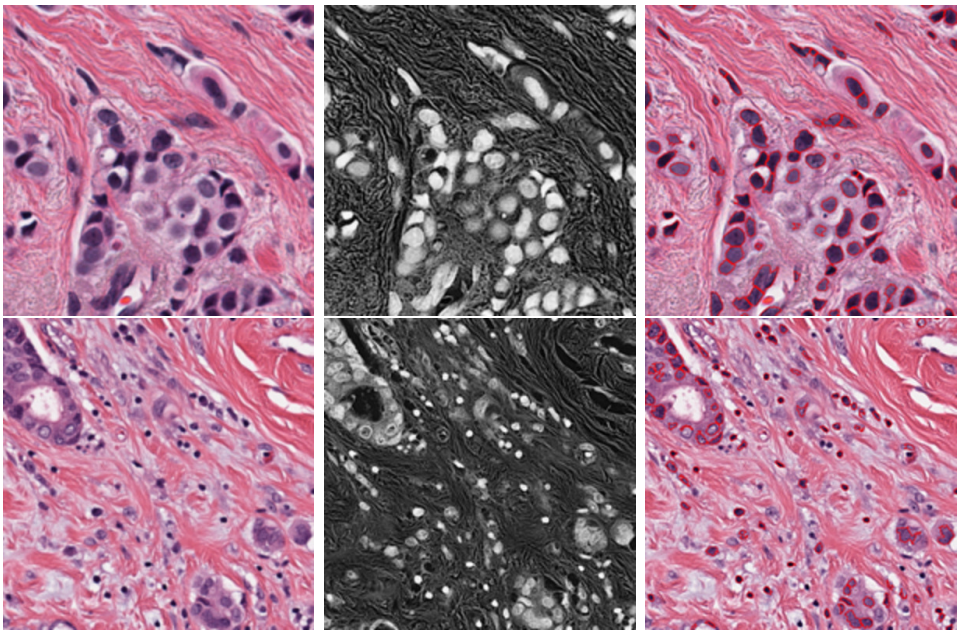
(a) Blur (b) Fold and Tear

(c) Annotations (d) Cracks

Figure 2.1: Examples of common artifacts in H&E images can be problematic for automated image analysis and lead to erroneous nuclear and cellular segmentation and features.

or unmixing, background removal, and artifact removal.

Stain unmixing, which quantifies the intensity of hematoxylin and eosin stains separately, is often used before processing, since the hematoxylin stain is especially useful for detecting and segmenting nuclei, as seen in Fig. 2.2. Stain normalization [29], which not only unmixes the stains, but also attempts to remove uninformative stain variation by normalizing to a reference stain, may also be used, and can be particularly helpful when working with large datasets from many tissue source sites (TSS), such as TCGA. With such datasets, "batch effect" issues often arise due to the variation across TSSs in instrumentation and data acquisition, which can significantly bias analysis [30]. Two example normalized patches from TCGA breast cancer patients using a popular recent method [29] are shown in Fig. 2.3. Still some research has cautioned against stain normalization [10], since not all stain variation is uninformative and since it can introduce yet other problematic



(a) H&E (b) Hematoxylin (c) Nuclei

Figure 2.2: Segmentation of nuclei in two small patches of WSIs. (a) The two patches. (b) The unmixed hematoxylin stain. (c) The segmented nuclei, bounded in red, by Otsu's thresholding algorithm. Results were obtained by CellProfiler software [17].

artifacts.

Methods for removing artifacts, or other uninformative regions of slides, such as background pixels, have also been proposed. Tissue folds are particularly troublesome for WSIs and methods for removing them have received attention in the literature [31].

In the proposed pipeline, both stain normalization and stain unmixing [29] are considered. No automated method for removing artifacts was used, rather; regions of WSIs were selected manually to ensure reliable analysis.

2.2 Cell and Nuclei Segmentation

Once the images have been refined by pre-processing, a common next step is segmentation of nuclei and their corresponding cells. The proposed algorithms for segmenting nuclei in H&E in the literature are numerous [27]. Watershed [32], active contour, and clustering-based methods have been proposed, but often are not used in the literature of H&E studies. This may be

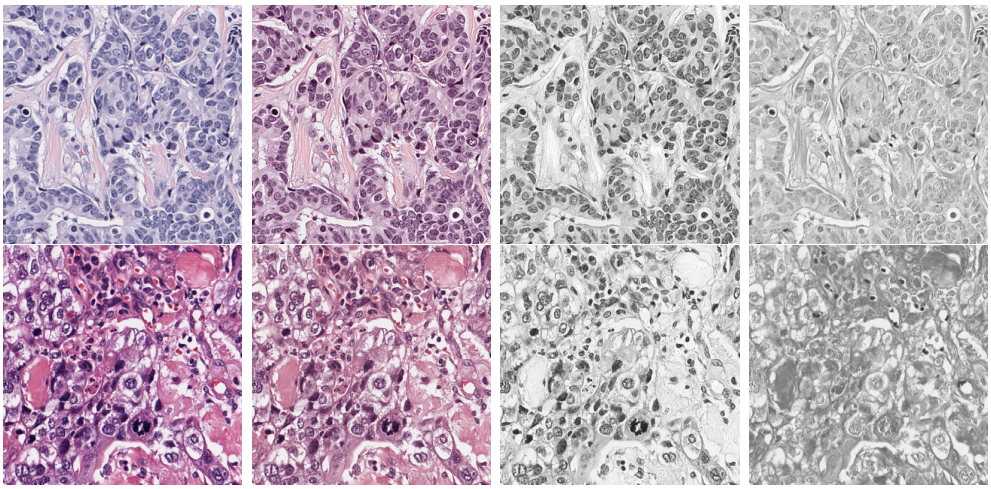
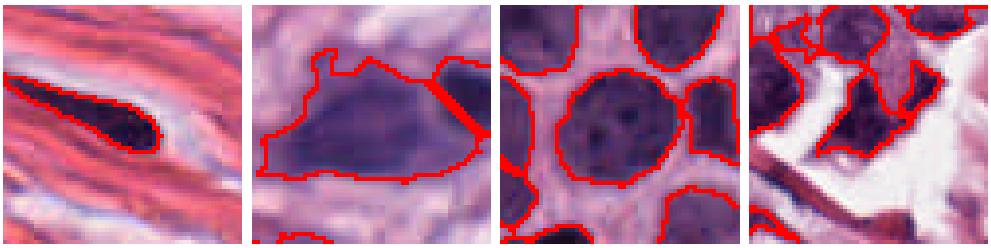


Figure 2.3: Stain normalized applied to H&E images of two TCGA breast cancer patients using the algorithm of [29]. Each row corresponds to a patient. Original H&E images (first column); normalized images (second column); hematoxylin channel (third column); eosin channel (fourth column). Dark pixels in the separated channels indicate a higher concentration of stain and light pixels a lower concentration.

due to the difficulty of tuning these algorithms for datasets with a high degree of variation across samples. A common method is Otsu's method [33], which assumes that the pixel intensities follow a bimodal or trimodal distribution and selects a threshold value that optimally separates these distributions. Often, simple thresholding methods, such as Otsu's method, are employed, which are straightforward, but also struggle to account for variation of intensity due to staining and sample thickness. Nuclear segmentation methods, and the CNN used in the proposed pipeline, will be described in more detail in the next chapter.

Segmenting cells is an ill-defined problem, since cell boundaries are often not visible from H&E staining. In most proposed methods, nuclei are detected first and segmented, and their positions are then provided to an algorithm to segment cells and classify them, or cellular segmentation is ignored altogether [5]. Cellular boundaries can be defined by Voronoi partitions [34], superpixels, or simply by neighboring pixels of the nuclei within a certain distance [12, 35, 14]. In other works, cells are segmented first and then nuclei are located within the cells. Superpixel segmentation [36] can also be used to define cellular boundaries first [19]. Our pipeline uses Otsu's thresholding to define the cytoplasmic region of the cell surrounding each nucleus, but limits



(a) Nucleus 1: Gabor = 16.97 Area = 339 Eccentricity = 0.98 AveInt = 0.88 StdInt = 0.079

(b) Nucleus 2: Gabor = 3.42 Area = 1263 Eccentricity = 0.80 AveInt = 0.77 StdInt = 0.086

(c) Nucleus 3: Gabor = 2.81 Area = 756 Eccentricity = 0.54 AveInt = 0.83 StdInt = 0.077

(d) Nucleus 4: Gabor = 11.21 Area = 294 Eccentricity = 0.60 AveInt = 0.83 StdInt = 0.099

Figure 2.4: Examples of segmented nuclei from TCGA breast cancer patients using Otsu's method in CellProfiler, along with several of their extracted features. Stromal nuclei (Nucleus 1) are long, thin, and surrounded by pink stromal tissue, and are discriminated by eccentricity measures. Epithelial cells are generally rounder and surrounded by purple cytoplasmic regions. Lymphocytes (not shown) are smaller, circular, and dark.

valid regions to be within 15 pixels of the nucleus.

2.3 Cellular and Nuclear Features

The primary statistics, or features, used in H&E analysis are those of cells and their nuclei, since pathologists' experience has deemed these features to be highly indicative of outlook. Generally, such features fall under the categories of shape, texture, and color. A comprehensive list of definitions of feature transforms commonly applied to H&E images can be found in [37].

Nuclear features in particular have been considered in computational pathology. The imaging-genomic works of Cooper and Kang *et al.* [38, 12, 13, 35, 14] performed their analyses using only such nuclear features and features of the surrounding cytoplasm. Often, these features are simple descriptors, such as the statistics of the intensity of hematoxylin stain in the nucleus and rotation invariant shape descriptors, such as the area of the nucleus and the length of its major and minor axes. Image texture can be captured by Gabor filters, which are oriented Gaussian filters that have been modulated by an oriented

sinusoidal. Images that have strongly oriented color patterns will have a strong response to the Gabor filters oriented orthogonally to the direction of these variations. Examples of several differing nuclei and their corresponding features are shown in Fig. 2.4.

2.4 Cell and Nuclei Classification

Once the boundaries of cells or nuclei have been extracted, some researchers have classified cells, and even their nuclei, by type based on computed nuclear and cellular features. Beck *et al.* [19] were able to discover the importance of stromal cell features in prognosis because they had first classified cells as stromal, epithelial, or other. Nuclei within each cell type were then subclassified as typical or atypical depending upon pixel size and roundness. Cells were classified using L_1-logistic regression with a sparsity regularization parameter. Training the classifier required hand-labeled subsets of superpixels from 158 images, which required about an hour of work from a pathologist. Their algorithm produced an 89% accuracy on held-out data.

Yuan *et al.* [5] segmented images into cancer, lymphocyte, and stromal cells using known correlations between cell type and nuclear morphology. A support vector machine classifier trained by a pathologist was used for classification using features based upon the size and shape of the nuclei of the cell, knowing that malignant cells generally have large, round nuclei, whereas lymphocytes have generally small (<8 μm), dark nuclei and not much cytoplasm, and stromal cells can have spindle-shaped nuclei. After an initial classification, they used a spatial kernel smoothing technique [39] and a hierarchical multiresolution model using global features of the tumor to refine the cell labeling, yielding a classification accuracy of 90.1% using cross-validation.

The current proposed pipeline does not include a cell classification stage, in part because creating a reliable classifier is challenging without accurate segmentation. However, as more labeled data is gathered for the CNN proposed in the next chapter, incorporating such a classifier will become feasible. Nuclei and cell classes can also be learned in an unsupervised fashion using clustering, which is considered later in this thesis.

2.5 Global Features

The work of Beck *et al.* [19] demonstrated the importance of considering not only the features of individual cells and their nuclei, but also the spatial relationship between cells and nuclei. They found that the presence of stromal regions with many nuclei and stromal spindle nuclei bordering stromal round nuclei were both indicators of poor prognosis, which would not have been discovered without features capturing global structure of the tumor. Many metrics can be used to capture such global features, such as density and number of bordering pixels. Graphical model representations can also be used, as described in the review of Gurcan *et al.* [24]. The current pipeline does not incorporate spacial relationships, but adding them is ongoing work.

2.6 Slide Summarization

Once the features are extracted, a challenge is summarizing the features of all the nuclei and cells present in each image, which can be on the order of thousands for WSIs. The features from each cell and nuclei must be aggregated in a way that is invariant to translation and rotation, preserves as much information as possible, and is also succinct for purposes of computation. In several prominent works, this summarization is merely an average of nuclei an cell features across the entire image [35, 14, 28], or large partitions of the image, though some amend this summarization with the addition of second-order statistics of features. Our pipeline computes the mean, standard deviation, and percentiles of the features of each nucleus and cell across a subset of representative patches from the WSI to capture their distribution throughout the WSI.

2.7 Proposed H&E Analysis Pipeline

In summary, the proposed H&E image analysis pipeline transforms patches of a WSI into a single nuclear and cellular image feature descriptor, consisting of the statistics of texture, shape, color, and intensity features derived from its cellular and nuclear boundaries. The pipeline first unmixes the H&E stains, normalizes them [29], and then estimates binary nuclear segmentation

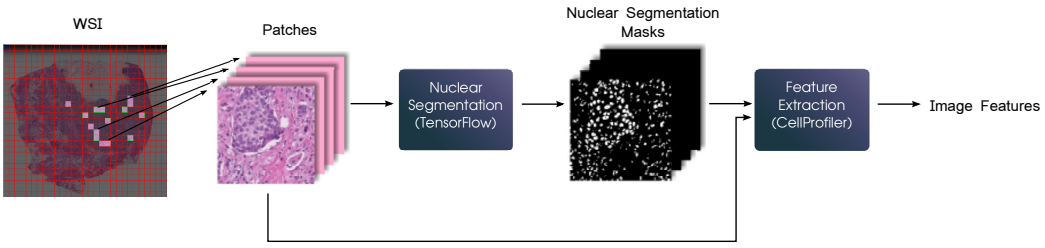


Figure 2.5: Combined nuclear segmentation and feature extraction pipeline for WSI patches.

masks for each patch via a CNN implemented in TensorFlow [40], which will be described in the following chapter. This CNN performs significantly better segmentation and detection of nuclei compared to thresholding methods, like those of [5] or CellProfiler. The segmentation masks and original H&E images are then fed to CellProfiler to refine them using spatial smoothing and to separate clumped nuclei, as shown in Fig. 2.5. CellProfiler then determines the boundaries of the cell corresponding to each nucleus using color thresholding, restricting the region to be within 15 pixels of the nucleus. Such a boundary still allows for meaningful cellular features. From the nuclear and cellular segmentation boundaries computed with our CNN and CellProfiler's refinement steps, CellProfiler extracts a total of 219 image features for each cell-nucleus pair, describing shape, color, intensity, and texture. To summarize the features of all nuclei and cells for a patient, the distribution, including the mean, standard deviation, and percentiles, of each feature is calculated, comprising a single image feature descriptor of 2409 features for the WSI. If multiple patches are selected from a WSI, then these statistics are calculated across all patches. In our experiments, up to 15 patches were manually selected for each WSI from representative regions to avoid contamination by artifacts and to make computation feasible, though any subset of patches could be use.

In the next chapter, the effectiveness of our proposed network over standard threshold-based methods, such as those used in CellProfiler, is demonstrated. Subsequent chapters include demonstrations of the efficacy of this pipeline on TCGA-BRCA images, revealing connections between image-based clusters and gene expression, with associations with patient outlook.

CHAPTER 3

CNN NUCLEAR DETECTION AND SEGMENTATION

3.1 Overview and Related Work

Fundamental to H&E analysis is reliable delineation of nuclear and cellular boundaries. Since nuclei receive more of the blue color of the hematoxylin stain and cytoplasm receives more of the pink color of the counter stain, eosin, color thresholding is commonly used for segmentation. There are many approaches to nuclear segmentation available in the literature [24, 27], but often in computational pathology studies, only simple methods are used. This is likely due to the difficulty in tuning parameters of more complex methods, the lack of available code and documentation for non-specialists, or lack of confidence in the robustness of these methods to variation in the H&E images, such as stain saturation and tissue thickness, that are inherent to the sample acquisition and preparation process. Certainly, there is a need for bridging the gap between advances in histopathological image processing and computational pathology research.

The most popular method for segmentation is a simple thresholding strategy, namely Otsu's thresholding [33], which is intuitive, straightforward to implement, and is available in software packages such as CellProfiler [17]. This method has been used in several prominent computational pathology studies [28, 5]. The image analysis tool CRImage [5], used for segmenting nuclei and classifying them as cancer, stroma, or lymphocytes, uses this strategy with an additional spatial smoothing step. Though this algorithm is attractively simple, it also struggles to handle varieties of staining and slide thickness and variation of texture among nuclei, even when applied adaptively. Especially in non-uniformly acquired data, such as that in TCGA, variation is high and segmentation is difficult.

Another approach seen in computational pathology is superpixel segmen-

tation [10], sometimes with a multi-resolution component [19]. This method first segments the image into superpixels, which can be considered to correspond to cells, and then nuclei are segmented from within these regions using color thresholding.

3.2 Convolutional Neural Networks

Recently, with the advent of the era of big data, neural networks have had a renewed interest in the image processing community, among others, under the banner of "deep learning," and have exhibited state-of-the-art performance on most machine learning tasks. Although neural networks were first proposed several decades ago, deep networks were not at first found to perform better than simple, shallow architectures, and these methods fell out of fashion. The primary obstacle in obtaining good performance is the challenge of optimizing the many parameters of a deep network. The initial strategy of random weight initialization with back-propagation often led to poor local minima that did not lead to good performance. It was not until recently [41], in 2006, that the idea of greedy, layer-wise pre-training of deep networks succeeded in reaching state-of-the-art performance on machine learning benchmarks.

Generally, a neural network is a directed network of layers of filtering, pooling, and an activation function, with a classification layer at the final output:

$$x_{l+1} = f\left(g_l\left(W_l^i x_l\right) + b_l^i\right), \tag{3.1}$$

where l indexes the layer, i indexes the filter W and offset b, g is the pooling or stride operator, and f is the activation function. A variety of choices exist for activation functions, such as the rectified linear unit, sigmoid, or hyperbolic tangent function; pooling methods, such as min, max, or average; stride levels; and classification layer, often the softmax function, along with the choice of number of layers and filters per layer.

A particularly powerful insight for image classification tasks is that images often consist of small, local patterns that are repeated across the image. Enforcing this insight in a neural network can be achieved by restricting the connections between layers to be of the form of local filters and applying the

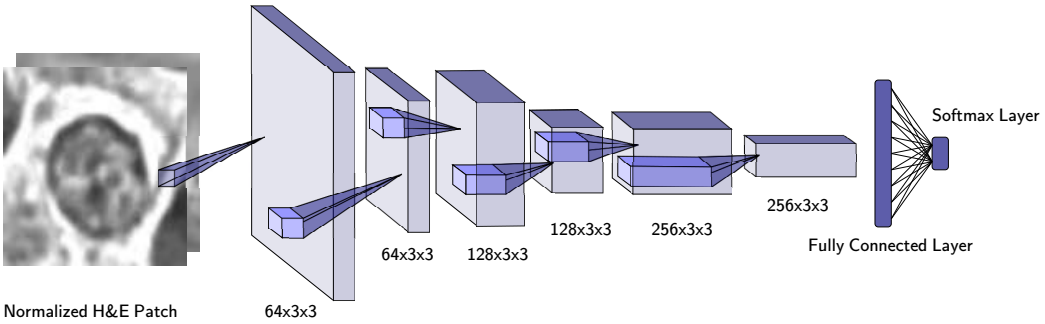


Figure 3.1: Proposed CNN architecture.

same filters across the entire image, a process known as convolution. These convolutional filters learn local features for a particular layer, and through the hierarchical structure of the network, effective means for combining the information from across disparate regions of the image are learned. Convolutional networks also have earlier beginnings [42], but have been applied more recently to image classification tasks, such as ImageNet [43], with remarkable success.

Recently, deep learning has been proposed for the task of nuclear detection [44, 45] and segmentation [46, 47] on H&E images, and has shown promising results. It has also been used for nuclear classification, such as mitosis counting [48], an important metric for prognosis. Especially as the collective set of publicly available H&E image data grows, neural networks will continue to become more relevant for analysis. Of these methods, the work of [47] is the closest comparable algorithm, since it was trained on breast images, and a working model is provided online, though it was trained specifically for estrogen receptor positive epithelial nuclei. Others perform only detection and not segmentation [44, 45] or did not provide a model or data for training a model that might be readily available to the research community [46]. Our network requires only TensorFlow and our trained model to run, with no parameters to tune, and was trained on patient data from TCGA-BRCA from a variety of BRCA types and tissue source sites, along with the data available from [47]

Our proposed network, diagrammed in Fig. 3.1, consists of six convolutional layers of $\{64, 64, 128, 128, 256, 256\}$ $3 \times 3 \times N$ filters, where N is the number of filters of the previous layer. Before being fed through the network, each patch is unmixed into its hematoxylin and eosin stains and normalized [29] to mitigate stain variation and to reduce the last dimension

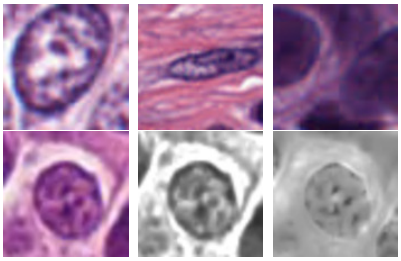


Figure 3.2: Example epithelial nucleus (top-left), stromal nucleus (top-middle) and non-nucleus (top-right) patches of 51 $\times$ 51 pixels, which are used as training for our CNN. Each patch, such as the example shown (bottom-left), is unmixed into its hematoxylin (bottom-middle) and eosin (bottom-right) stains and then copied at rotations of 90 degrees and horizontal and vertical flips to promote invariance to these operations.

N of the input layer filters from three to two. At every other layer, starting at the second layer, the convolution operator is applied at a stride of two, which is equivalent to downsampling the input layer by two in both spatial dimensions. The final two layers are a fully-connected layer of 50 nodes and a softmax output layer of two nodes. The CNN operates on inputs of 51 $\times$ 51 pixel patches, producing a binary label for the center pixel of each patch indicating whether it belongs to a nucleus or not. To produce a mask for the entire image, each patch in the image is processed by the CNN.

For training of our CNN, we manually labeled and extracted a dataset of several hundred sample patches of nuclei and non-nuclei from a set of 68 TCGA-BRCA patients, comprising 32,174 patches and representing a variety of TSSs. The pre-processing of each patch consists of normalizing the stain [29], separating the hematoxylin and eosin images, and then generating rotations at 90 degree increments, as well as horizontal and vertical flips of the images, to promote invariance to such manipulations, which naturally arise in H&E images. Several example patches are shown in Fig. 3.2, along with an example patch unmixed into hematoxylin and eosin stains.

Once the initial binary segmentation mask for each WSI patch is generated by the CNN, the mask along with the corresponding H&E image are passed to CellProfiler to be further enhanced by smoothing and to separate clumped nuclei. Several example WSI patches, and the resulting binary masks and overlayed, refined boundaries of nuclei and cells, from TCGA-BRCA patients are shown in Fig. 3.3.

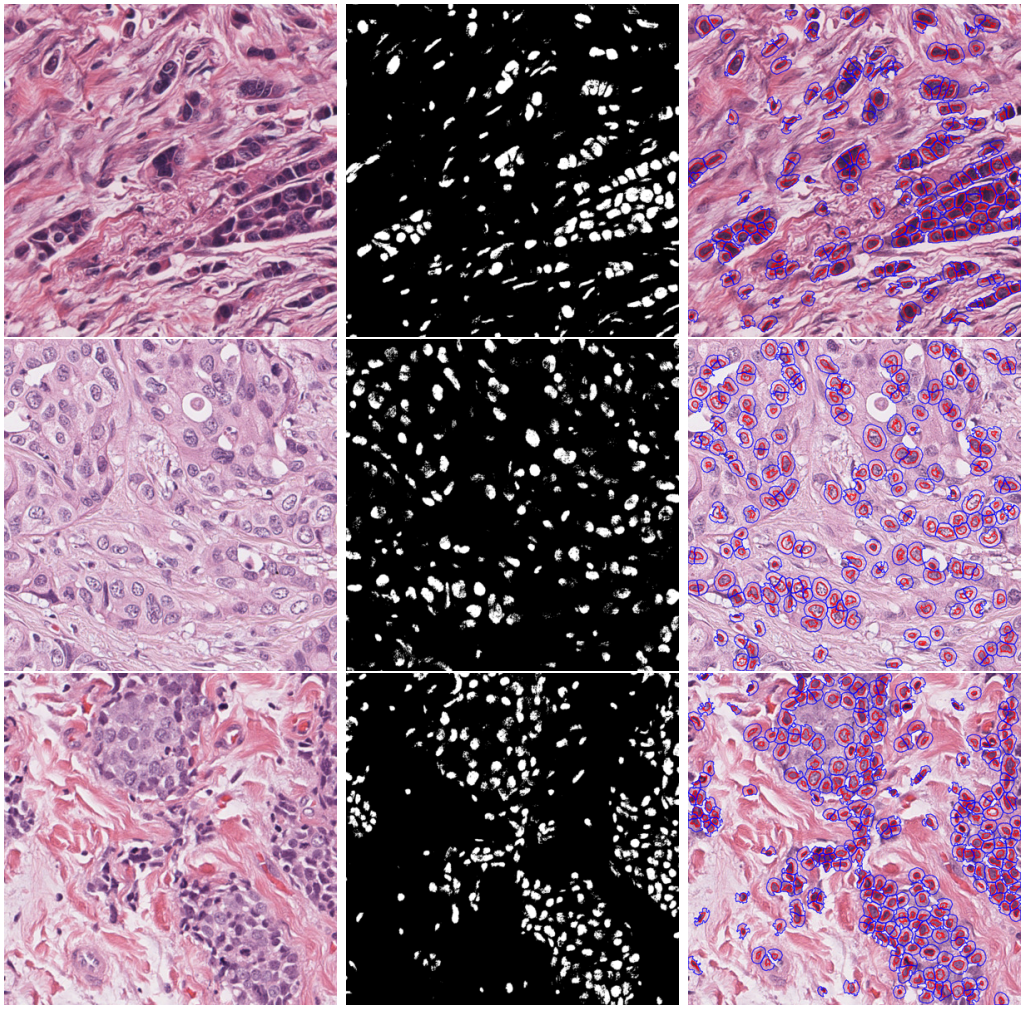


Figure 3.3: Example TCGA-BRCA WSI patches (left) are first segmented by the proposed CNN, yielding binary nuclear masks (middle), and then the mask, along with the original patch, are fed to CellProfiler to refine the masks and overlay nuclear and cellular boundaries (right). Each row corresponds to a different sample patch.

3.3 Evaluation

We evaluated our CNN on nuclei detection and segmentation tasks to compare with other similar available tools. To our knowledge, the only publicly available dataset of breast cancer H&E histology images with ground truth labeled nuclei, other than training data provided by [47], is the University of California, Santa Barbara (UCSB) biosegmentation benchmark [49]. The labeled images for this dataset are pixel-wise binary masks of nuclei pixels. Boundaries between touching nuclei were not delineated on the masks; so in

Table 3.1: Comparative detection accuracy of our CNN, our CellProfiler pipeline, and the network of Janowczyk *et al.,* 2016 [47] on UCSB breast cancer H&E images.

Distance	Algorithm	Precision	Recall	F1-Score
$t = 15$	CNN	0.841	0.910	0.874
	Janowczyk *et al.*, 2016	0.855	0.876	0.866
	CellProfiler	0.915	0.760	0.831
$t = 12$	CNN	0.830	0.875	0.852
	Janowczyk *et al.*, 2016	0.850	0.850	0.850
	CellProfiler	0.905	0.712	0.800
$t = 10$	CNN	0.820	0.857	0.838
	Janowczyk *et al.*, 2016	0.844	0.832	0.838
	CellProfiler	0.890	0.676	0.768

order to ensure objectivity and reproducibility, we inferred these boundaries automatically using CellProfiler's nuclei separation tool. These images were captured at a lower magnification than 40×, the magnification on which both our network and the network of [47] were trained, so the images were resized by a factor of 2 to approximate 40× magnification. We refer to the resulting separated nuclei masks as the gold standard.

3.3.1 Detection

Precision, recall, and the F1-score for evaluating detection performance are computed as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

where TP is a true positive, FP is a false positive, and FN is a false negative. In our evaluation, a nucleus is considered a true positive if its center is within a specified distance, in terms of pixels, of the center of a nucleus in the gold standard mask. When a gold standard nucleus matches multiple predicted nuclei, its closest match is chosen.

Table 3.2: Detection accuracy of CNN on colon H&E images [44].

Algorithm	TP	FP	FN	Precision	Recall	F1-Score
CNN	21933	11503	7383	0.66	0.75	0.70

The scores for our network, our CellProfiler pipeline, and [47] with varying values of the acceptable distance threshold are shown in Table 3.1. The network of [47] produces a probability, which allows for user tuning by varying the threshold to be applied to make a binary decision. We evaluated thresholds ranging from 0 to 0.92 by increments of 0.04 and reported the the threshold with the best F1-score. Our algorithm performs comparably with that of [47], but requires no parameter tuning, which could not be performed so precisely on datasets for which there is no gold standard reference and high variation of staining. More pertinent is that our algorithm shows a marked improvement over the thresholding technique of CellProfiler.

To evaluate the versatility of the proposed network to generalize to other cancer types, it was also evaluated on the dataset of H&E images of colon cancer from [44]. Precision, recall, and F1-scores are shown in Table 3.2. The scores of our algorithm fell short of the reported results of an F1-score of 0.802. Inspecting the resulting images, shown in Fig. 3.4, revealed that the proposed network struggles particularly on faint, thin stromal nuclei. Comparing qualitatively the results of both networks, shown in Fig. 3.5, revealed that in some cases, the networks perform nearly the same, with similar false positives and negatives.

3.3.2 Segmentation

A wealth of metrics exists for evaluating image segmentation, with each metric capturing different aspects of performance. The recent work of [46] chose the the Dice similarity coefficient (DSC), the Hausdorff distance (HD), and the mean absolute distance (MAD), which we adopted for consistency and comparison. For a given nucleus, let Ω_{gs} denote its gold standard segmented region in the image and $\hat{\Omega}$ the estimated region. The Dice similarity coeffi-

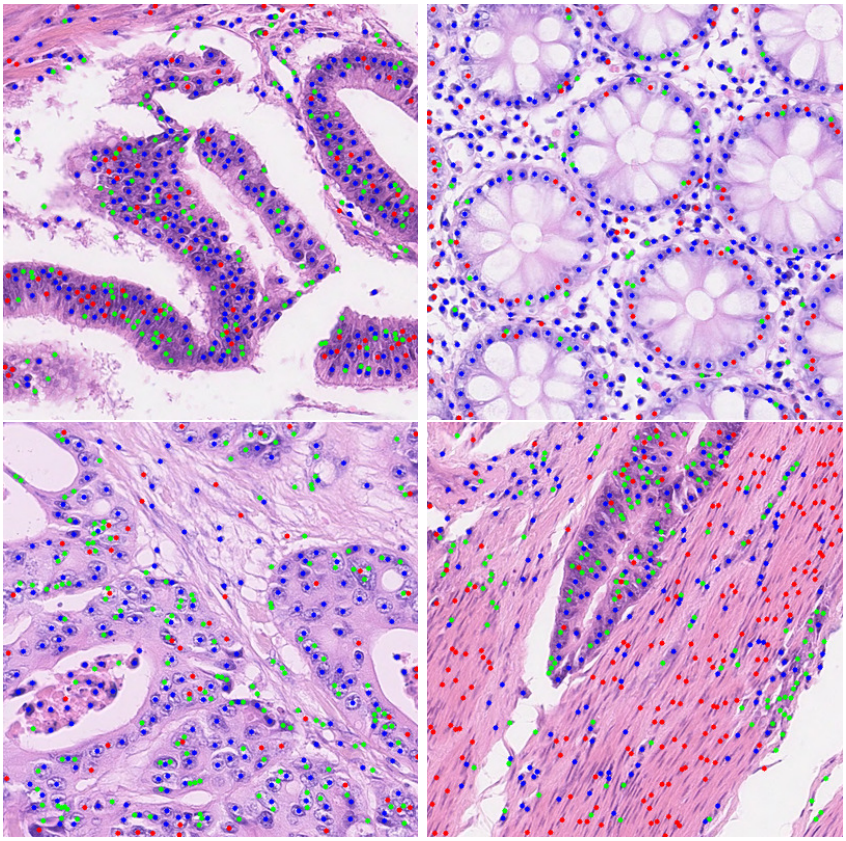


Figure 3.4: Detection results of the CNN on colon H&E images from [44]. True positive nuclei locations are shown in blue, false positives in green, and false negatives in red. Each image is a different colon H&E sample from the dataset. The network performs well on these images, despite it not being trained on colon tissue. In the bottom left image, it suffers from over-segmentation of nuclei, indicated by the numerous green dots. In the bottom right image, the presence of many red dots in the stromal tissue indicate that the network struggles systematically to detect the thin, long stromal nuclei.

Table 3.3: Comparative segmentation accuracy of our CNN, our CellProfiler pipeline, and the network of [47] UCSB breast cancer H&E images.

Distance	Algorithm	DSC Mean	DSC Median	DSC STD	HD Mean	HD Median	HD STD	MAD Mean	MAD Median	MAD STD
$t = 15$	CNN	0.72	0.79	0.20	4.38	3.16	3.89	1.85	1.31	1.42
	Janowczyk *et al.*	0.75	0.81	0.16	4.22	2.83	4.03	1.75	1.29	1.20
	CellProfiler	0.76	0.83	0.18	4.62	2.83	4.70	1.84	1.24	1.64
$t = 12$	CNN	0.72	0.80	0.20	4.24	3.00	3.83	1.82	1.26	1.43
	Janowczyk *et al.*	0.75	0.81	0.16	4.22	2.83	4.03	1.75	1.29	1.20
	CellProfiler	0.76	0.82	0.18	4.62	2.83	4.70	1.84	1.24	1.64
$t = 10$	CNN	0.69	0.78	0.25	4.77	3.16	4.43	2.15	1.37	1.97
	Janowczyk *et al.*	0.72	0.80	0.21	4.69	3.00	4.52	2.01	1.35	1.68
	CellProfiler	0.70	0.796	0.25	5.68	3.16	5.93	2.41	1.46	2.32

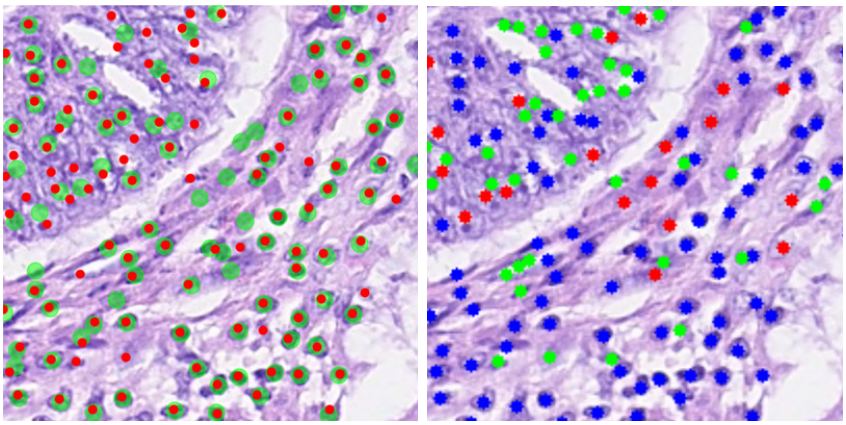


Figure 3.5: Comparison of detection results of [44] (left) and the proposed CNN (right) on an example colon H&E image. For the CNN, true positive nuclei locations are shown as blue dots, false positives as green dots, and false negatives as red dots. For [44], detected nuclei locations are shown as red dots and locations acceptably close to the ground truth are encapsulated by larger green circles. In several places, both algorithms produced false positives or negatives at the same locations, suggesting that our CNN is learning to detect similar patterns, despite being trained on different tissue.

cient measures the relative overlap of the regions:

$$\mathrm{DSC} = 2\frac{\left|\hat{\Omega} \cap \Omega_{gs}\right|}{\left|\hat{\Omega}\right| + |\Omega_{gs}|}. \tag{3.5}$$

The Hausdorff distance is commonly used to compare surfaces or boundaries. It measures the maximum deviation between the two boundaries:

$$\mathrm{HD} = \max\left\{\sup d(v_{gs}(s), \hat{v}), \sup d(\hat{v}(s), v_{gs})\right\}, \tag{3.6}$$

where $d()$ measures the Euclidean distance between the boundaries and $v_{gs}(s)$ and $\hat{v}(s)$ are the boundaries of the gold standard and estimated segmented regions, respectively. The mean absolute distance between the boundaries is computed by

$$\mathrm{MAD} = \frac{\int d(v_{gs}(s), \hat{v})|\hat{v}'(s)|ds}{2|\hat{v}(s)|} + \frac{\int d(\hat{v}(s), v_{gs})|v'_{gs}(s)|ds}{2|v_{gs}(s)|}. \tag{3.7}$$

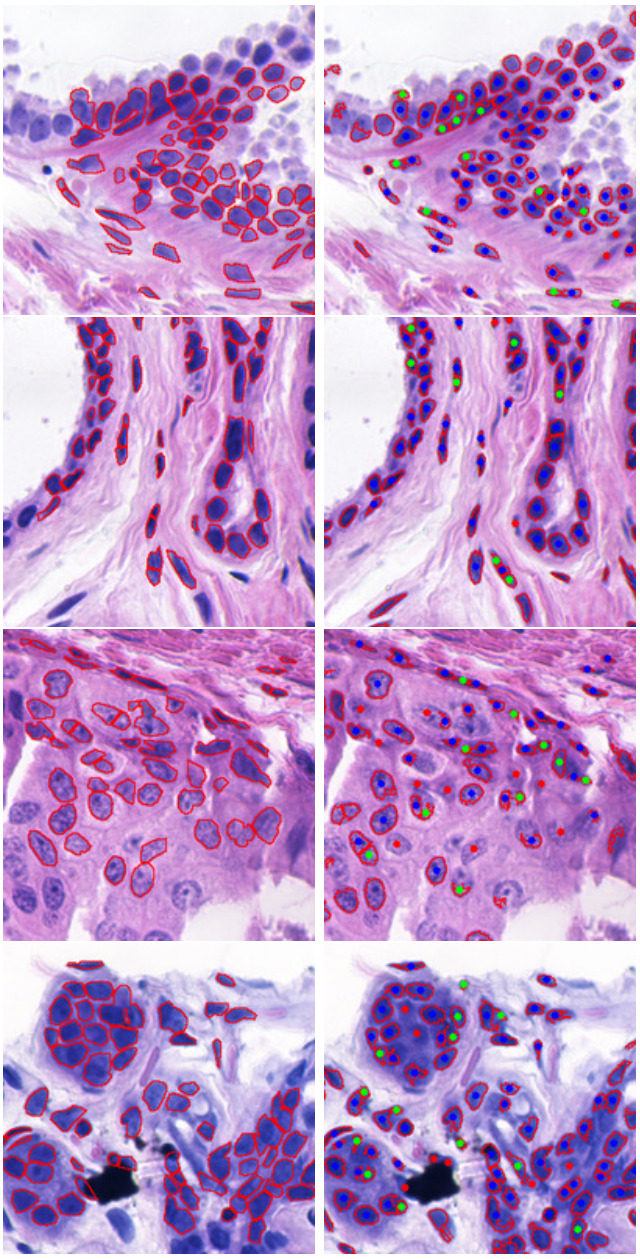


Figure 3.6: Segmentation results of our CNN (right column), and the gold standard (left column) from UCSB H&E breast cancer images. Each row corresponds to a different sample in the dataset. Blue dots in the right images indicate true positive nuclei locations generated by our CNN, green dots false positives, and red dots false negatives. Nuclei boundaries from our CNN are delineated in red. Our CNN performs well on the bottom and top two images, but misses several nuclei in the image of the third row, likely due to its fainter color, the lack of texture of the nuclei, and possibly a resolution mismatch.

Since these metrics can only be applied to true positive nuclei, they do not capture the effects of false negatives and false positives, so a desired balance must be considered when comparing algorithms. A comparison of the results for each of these metrics is shown in Table 3.3. Again, our CNN and the network of [47] perform similarly, though ours slightly worse. CellProfiler performs slightly better in terms of DSC, but worse in both HD and MAD, despite its having a significantly higher precision score, which indicates that it is more conservative in what it detects as nuclei. Several example gold standard images and their segmentation by of our CNN are shown in Fig. 3.6. The algorithm is able to detect and segment both stromal and epithelial nuclei, but struggles with nuclei with fainter hematoxylin stain. This could be a consequence of the difference in resolution between the 20× UCSB images and the 40× TCGA images on which it was trained.

Datasets such as TCGA pose a much greater difficulty for segmentation since the acquisition procedures across the various TSSs are less controlled and prone to introduce significant variation. A prominent advantage of our network is that it has been trained on WSIs from TCGA-BRCA patients, increasing its robustness to these variations. We chose a small subset of WSI patches with varying slide quality and hue from TCGA-BRCA patients on which to qualitatively evaluate our CNN, with CellProfiler's refinement steps, and compare with CellProfiler's built-in threshold-based segmentation. The overall segmentation pipeline in CellProfiler consisted of adaptive thresholding to remove white background pixels, adaptive three-class thresholding to segment nuclei, declumping of nuclei, and a filtering stage to remove segmented objects outside of a specified range. Finding limits of the respective thresholds that yielded good performance across all images was difficult. We set the white pixel background threshold to be from 0 to 0.3 and the adaptive three-class threshold for nuclei segmentation to be from 0.4 to 1. The yielded nuclear and cellular boundaries of both methods for three images from the test set are shown in Fig. 3.7. Overall, our CNN was able to perform much better on the set of images and required no parameter tuning, unlike CellProfiler.

The proposed network was also compared to the thresholding method of CRImage [5]. Several example images of lung adenocarcinoma from TCGA are shown in Fig. 3.8. On an example with stark contrast between the color content of stromal tissue and nuclei, CRImage performs similarly to the

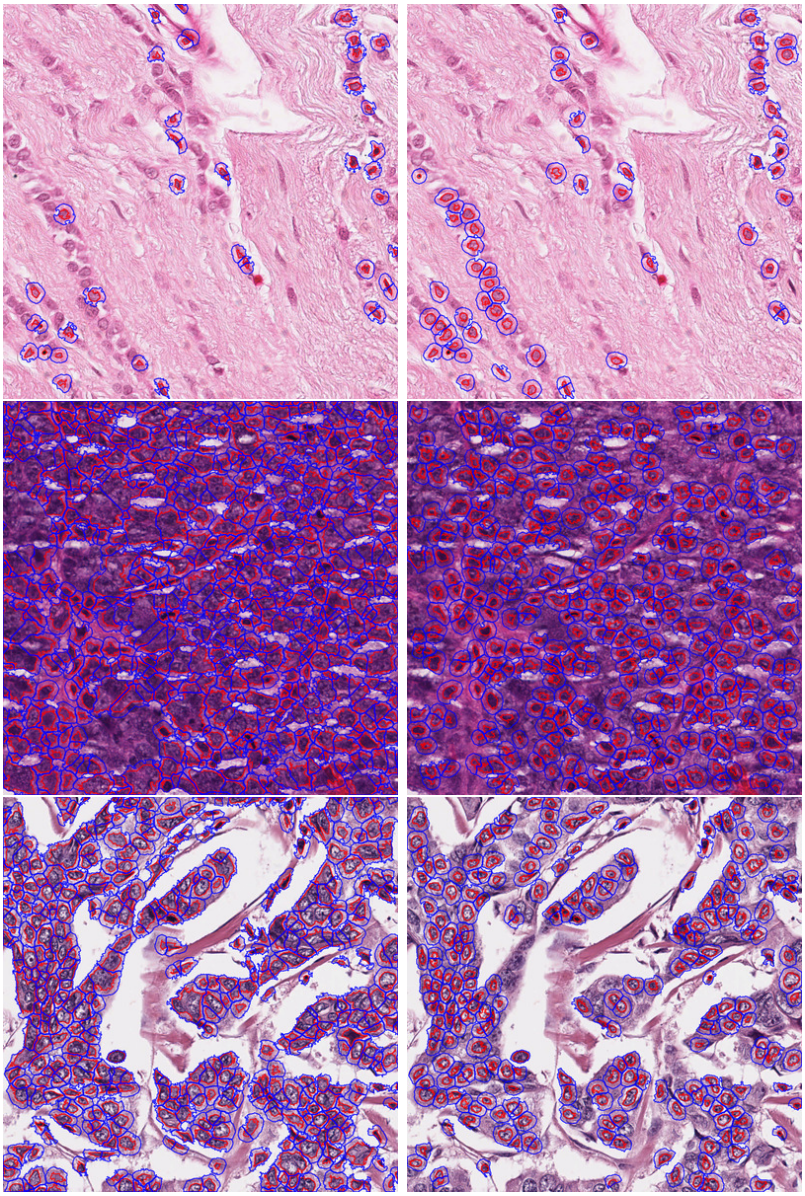


Figure 3.7: Cell and nucleus segmentation results of our CNN (right column) and CellProfiler (right column) on three diagnostic WSI patches of the TCGA-BRCA dataset. Nuclear boundaries are drawn in red and cellular boundaries in blue. Our CNN is able to robustly detect and segment nuclei despite variation due to staining and slice thickness, whereas the thresholding approach of CellProfiler is not robust to such variation. Increasing the threshold improved performance on darker images, such as the bottom two rows, but at the cost of missing most nuclei in lighter images, such as the top row. Our CNN was able to detect and segment nuclei well despite the stark differences in intensity and hue. In particular, it was better able to avoid clumping large nuclei together, as seen in the comparison of the bottom two rows.

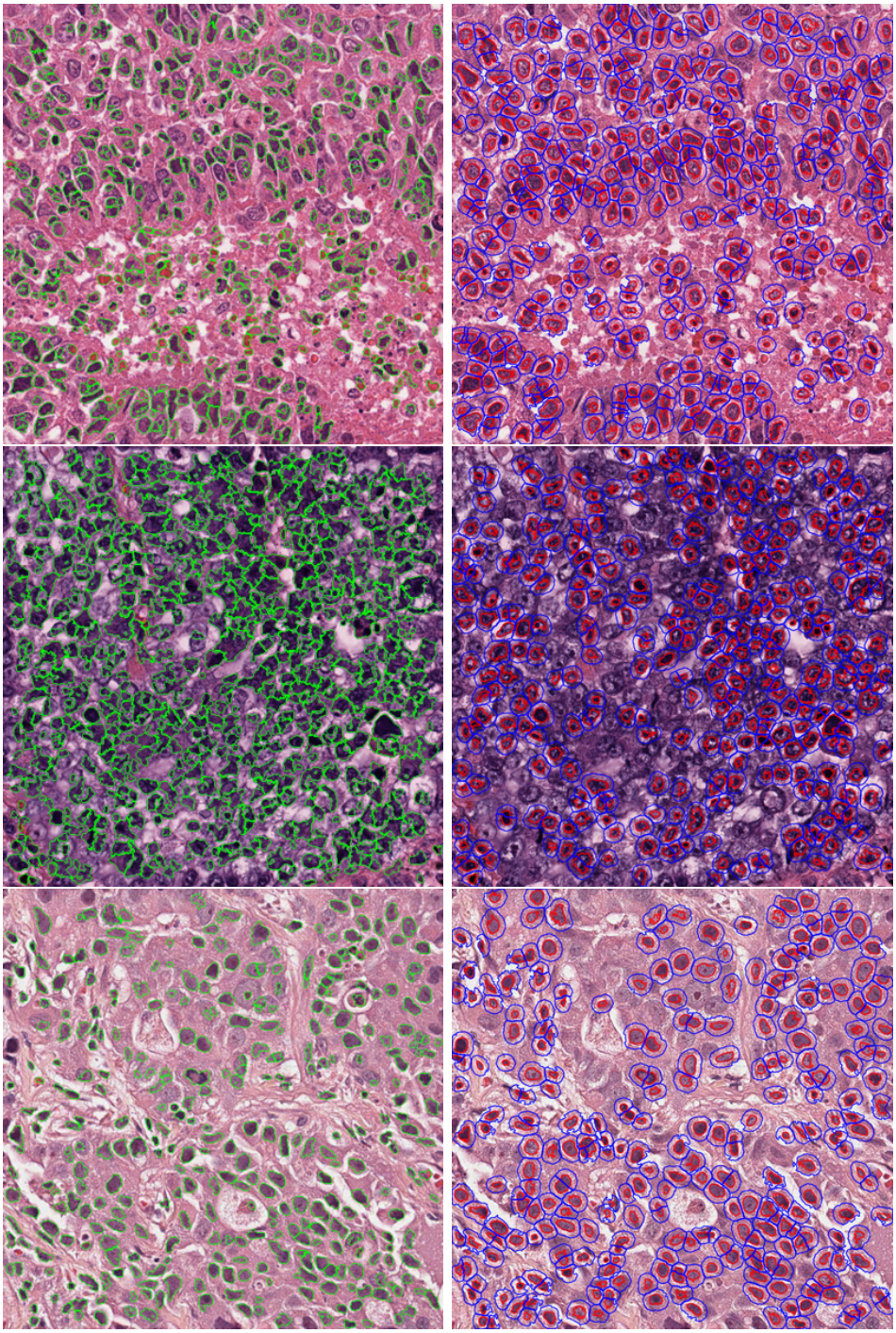


Figure 3.8: Segmentation results of CRImage [5] (left column) and the proposed CNN (right column) on TCGA-LUAD diagnostic H&E image patches. Each row shows the results of the two algorithms on a different patch. The two algorithms perform similarly on the patch of the bottom row, but the CNN performs noticeably better on the patch of the top and middle rows, which are exemplary of the more challenging cases existing in such datasets.

CNN. However, for more challenging cases, CRImage produces many poor boundaries and false positives, whereas the CNN performs much better, and though it misses some large epithelial nuclei, the nuclei it does detect and segment are reliable.

3.4 Codebook Quantization of TCGA-BRCA Nuclei

Ultimately, the goal of nuclear segmentation is to provide an accurate and reliable boundary for subsequent feature extraction. If the segmentation is reliable, and the extracted features capture salient characteristics of nuclei, then these features should be discriminative of the various types of nuclei observed in breast cancer. Additionally, nuclei that have similar feature vectors, in terms of the Euclidean distance between them, should be perceptually similar.

Inspecting various nuclei and comparing their visual similarity to the Euclidean distance between their corresponding feature vectors would reveal the validity of these suppositions. A yet more informative approach is to first cluster the nuclei into groups with similar features and then inspect the visual similarity of representative nuclei of these groups. This procedure is known as *codebook quantization*, and is popular in image processing for summarizing data according to their distribution [50].

To perform this investigation, a subset of 500 BRCA patients from TCGA were processed by the pipeline, using the CNN to first generate the segmentation masks, and 100 nuclei with corresponding cellular features were chosen at random from each patient, constituting a dataset of 50,000 nuclei. For each nucleus and corresponding cell, 219 features describing shape, texture, and color were extracted. Using a Gaussian mixture model, the space of the features of all of these nuclei was summarized by a set of 200 clusters. In the context of codebook quantization, the vectors representing the center of each of these clusters are called *codewords*, and the set of these codewords composes the codebook.

To display these codewords in a way that conveys their relative distance in the feature space, they were first ordered according to relative distance. This ordering was derived by applying a greedy traveling salesman algorithm to the fully connected graph constructed from the cluster centers in

the 219-dimensional feature space, where each node is a cluster center and the weight of each edge is the Euclidean distance between the centers. The traveling salesman algorithm searches for a path through the graph that traverses every node with minimal total edge weights. Once this ordering was determined, each codeword was represented by the nucleus with the closest feature vector. These representative nuclei and their ordering are shown in Fig. 3.9. Traversing the nuclei by column and then by row, starting at the upper left and ending in the lower right, traces the path through the graph that was retrieved by the traveling salesman algorithm.

From this visualization, it is apparent that neighboring codewords indeed correspond to perceptually similar nuclei, and that the extracted nuclear and cellular features are capturing informative characteristics of shape, size, and texture. Notably, the nuclei are not ordered according to uninformative similarities, such as stain intensity, slice thickness, or other artifacts of the image acquisition process. There is also a diverse set of nuclei represented: some nuclei are quite small and dark, like lymphocytes; some are long, thin, and surrounded by pink stromal tissue, like stromal cells; and a variety of epithelial cells, some highly textured, some homogeneous, and ranging in size, are present. Additionally, nuclei of similar types are near one another in the ordering, indicating that these features are indeed discriminative of these various types.

As mentioned in the previous chapter, a stage that could be added to the proposed H&E analysis pipeline is cell and nucleus classification. Using labels to represent nuclei instead of a feature vector has many benefits. In particular, labels are more easily interpreted than a high-dimensional vector of features and they are amenable to spatial reasoning using graphical models, where each nucleus could be represented by a node with its label. However, acquiring labeled nuclei, especially with the diversity need to capture the variety within each type, is laborious and requires the costly time of a trained pathologist. Instead of using labeled examples to learn a classifier, such codewords could be considered unsupervised nuclei classes, and each nucleus could be labeled, or quanitized, according to its most similar codeword.

In order for classifiers, even unsupervised ones, to work effectively, they must be trained with a large set of examples that convey the natural variation present in each class. The TCGA dataset is a tremendous resource for learning such classifiers because of the quantity and diversity of nuclei

present in its hundreds of WSIs. Leveraging the richness of this dataset for codebook generation, such as the one presented, and developing new metrics that make use of these labels are promising directions for future cancer imaging research.

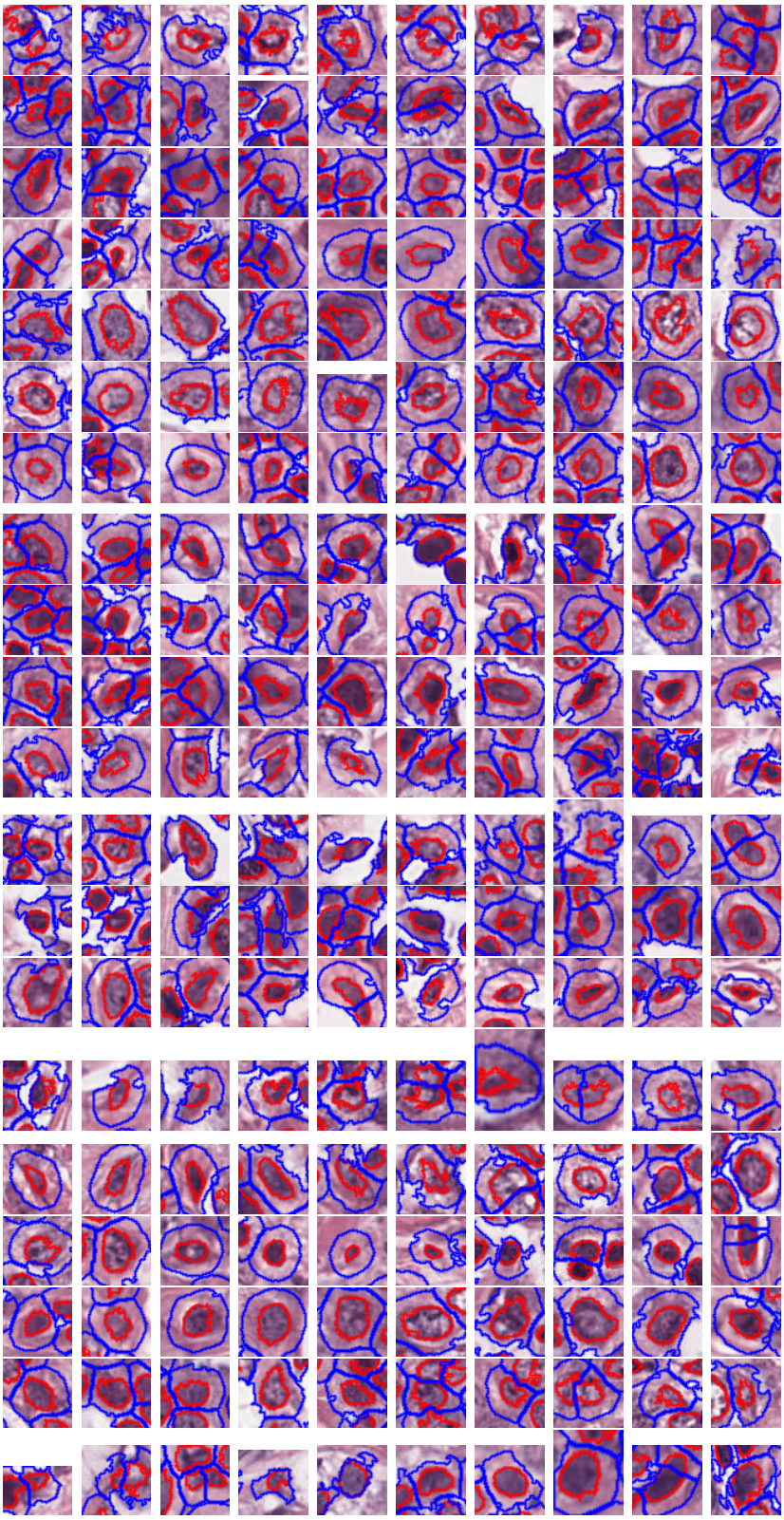


Figure 3.9: Representative nuclei from 200 codewords generated by GMM clustering nuclei features from 500 TCGA-BRCA patients.

CHAPTER 4

GENOMIC INTEGRATION

Genomic data consists of a variety of categories of measurements, including the presence of mutations, copy number variation, methylation, and mRNA expression. Though correlated, these different measurements provide unique insights into cellular function and the processes of gene transcription, regulation, and protein expression. For most commonly used methods of extracting these measurements from a section of tumor, DNA and RNA are aggregated from the entire section [51], and spatial information is lost. Another challenge of working with genomic data is that it falls in the "small n, large p" domain of statistical analysis; that is, in most studies, only a modest number n of samples (patients) is available, usually on the order of tens or hundreds, but the dimensionality p of each sample (number of genes) is vastly larger, with nearly 20,000 genes in the human genome.

A statistical challenge in this domain is inferring meaningful associations that are not due solely to noise or other sources of variation, such as those that may arise in the acquisition process or natural variation in the genome from patient to patient [52, 30]. Genome-wide association studies are the simplest approaches to significance testing, but are highly susceptible to discovering false associations. Statistical tools, such as significance analysis of microarrays (SAM) [53] in the case of gene expression, attempt to account for this possible trapping. Adding to the challenge, like H&E image data, gene expression is also subject to batch effects from different instruments or studies [30]. Additionally, it is known that genes can regulate and drive one another in highly complex pathways, which must be accounted for in the inference formulation. Associations must be considered not just for single genes, but for appropriate groupings of genes that act together.

This thesis primarily considers gene expression data, since it is most directly connected to the features observable by histopathologic imaging, along with a variety of techniques, leveraging understanding of biological pathways,

for reducing the dimensionality of expression data to a tractable level for inference. A number of aims can then be considered, including cancer subtyping via clustering, computational prognosis, correction of genomic data, and identification of imaging phenotypes to act as surrogates for genomic features. This thesis shows the efficacy of the previously described H&E analysis pipeline for discovering meaningful groupings of cancer patients related to outlook and investigates the genomic markers that are associated with these image-based groupings.

4.1 Genomic Data Dimension Reduction

To increase the statistical power of association tests, it can be helpful to reduce the dimensionality of genomic data, either through modeling that makes simplifying assumptions about the interactions of genes or their connections, or by leveraging the knowledge of cellular processes discovered by other studies. Several such approaches are reviewed here before considering the integration of genomic data with imaging.

4.1.1 Knowledge-Driven Gene Selection

One approach is to restrict the set of genes considered using biologic understanding from other studies. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [54] is a knowledge base of gene functions and interactions in cellular processes, represented graphically by pathways. For pathways that are known or suspected to be implicated in cancer, the specific set of associated genes can be investigated. The Catalogue of Somatic Mutations in Cancer (COSMIC) [55] is a comprehensive knowledge base for somatic mutations that have been implicated in cancer. These mutations are referred to as *driver* mutations, as opposed to *passenger* mutations, since they are believed to be the true cause of specific cancers amongst many observed mutations. The COSMIC gene census gives a list of roughly 500 somatic mutations that are linked to particular types of cancer and describes the particular dysfunction caused by each mutation.

4.1.2 Molecular Subtype

An important discovery in genomics research was the categorization of five intrinsic molecular subtypes of breast cancer: luminal A and B, basal-like, HER2-enriched, and normal [2]. These subtypes were found to have significantly distinct survival outlooks, with the basal-like subtype performing particularly poorly. A subset of 50 genes, known as the PAM50 subset, were identified to be discriminative of these types through gene expression clustering [2], which could be investigated specifically for associations with image data. Additionally, image data may be able to reveal new insights into each of these subtypes, possibly even the existence of prognositically informative image-based subtypes within each molecular subtype.

4.1.3 Linear Unmixing

Linear unmixing is a technique for unmixing high-dimensional data when it is assumed that each sample is generated as a linear mixture of a small set of common, base samples. The linear mixture model was formulated originally for hyperspectral imaging used in remote sensing, but it was recently applied successfully to gene expression data. Since gene expression is acquired from an aggregate of cells, and the aggregate will consist mostly of similar types of cells, if cells of the same type have similar gene expression, then the linear mixture model would be appropriate. A particular model developed for gene expression unmixing is the unsupervised Bayesian linear unmixing (uBLU) model, which performed well on a recent viral challenge to study gene expression patterns of influenza [56]. An advantage of the linear mixture model is that the resulting endmembers yield an interpretable meaning, as compared to other statistical dimension reduction techniques, such as PCA.

The linear mixture model assumes each data vector $\mathbf{s}_i$, in this case the gene expression of a patient, for patient i is generated as a linear mixture, usually a convex combination, of a small set of base *factors* or *endmembers* $\{\mathbf{e}_k \in \mathbb{R}^p\}_{k=0}^K$:

$$\mathbf{s}_i = \sum_{k=0}^{K} \alpha_k^{(i)} \mathbf{e}_k, \tag{4.1}$$

where $\left\{\alpha_k^{(i)}\right\}_{k=0}^K$ are the mixing coefficients for patient i and obey the prop-

erty: $\sum_{k=0}^{K} \alpha_k^{(i)} \leq 1$ ($\sum_{k=0}^{K} \alpha_k^{(i)} = 1$ in the convex case). The endmembers may be known *a priori*, or they may need to be discovered. A variety of algorithms exists for generating endmembers and for determining the mixing coefficients [57]. A popular algorithm is N-FINDR [58], which approaches the unmixing problem from a geometric perspective, searching for K pure pixels that compose the largest simplex encompassing all spectra. An enhancement of this, the normal compositional model [59], allows for the modeling of variation of endmembers by a Gaussian, where $\mathbf{x}_k^{(i)} \sim \mathcal{N}(\mathbf{e}_k, \sigma_k)$, and each observed spectrum is a linear mixture of vectors $\mathbf{x}_k^{(i)}$ sampled from the distribution of each endmember:

$$\mathbf{s}_i = \sum_{k=0}^{K} \alpha_k^{(i)} \mathbf{x}_k^{(i)}. \tag{4.2}$$

The uBLU algorithm is a further development of this model specifically for gene expression data.

Applying uBLU on the KEGG Jak-STAT signaling pathway, a pathway of 147 genes that is known to affect cell growth and is referenced in subsequent results, yielded six optimal endmembers for the model. The discovered endmembers, along with an example reconstruction of the expression of a TCGA-BRCA patient, are shown in Fig. 4.1. The distribution of the squared reconstruction error of gene expression of all 1100 patients is shown as well. For the example reconstructed gene expression shown, the mixing coefficients for the six factors were $\{0.0075, 0.0078, 0.0237, 0.5146, 0.0424, 0.4040\}$, indicating that only two factors contributed significantly to the reconstruction. The original expression levels for the Jak-STAT pathway genes from 1100 TCGA-BRCA patients are shown in Fig. 4.2, along with the results of hierarchical clustering, showing clear trends in the expression across patients. Through linear mixing, these trends can be captured succinctly by an appropriate and easily interpretable model, and each patient can be represented by their corresponding endmember coefficients, which can then be used for subsequent analysis.

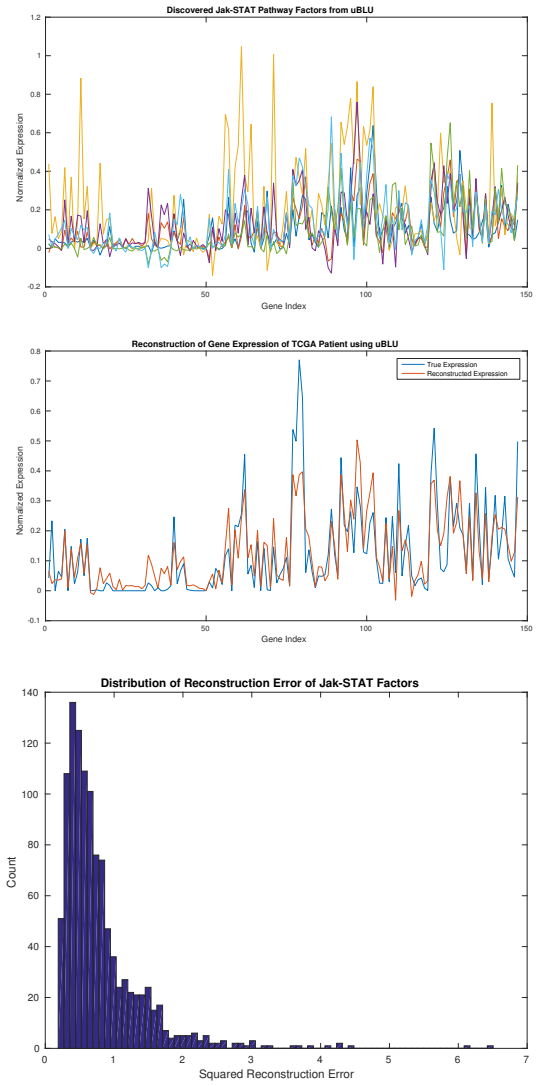


Figure 4.1: Discovered Jak-STAT factors using uBLU [56] for the 147 genes involved in the Jak-STAT signaling pathway in KEGG. The algorithm estimated that six factors (top) were adequate to reconstruct the expression data. The reconstructed expression for a sample TCGA-BRCA patient shown against the true expression (middle). The histogram of the squared reconstruction error of Jak-STAT genes for the six discovered factors for 1100 TCGA-BRCA patients (bottom) shows the effectiveness of the factors to represent the expression levels. The average squared error was 0.8011 and the average squared magnitude of the expressions of patients was 3.9486.

4.2 Data Pre-Processing

Before considering methods of genomic and image integration, the data used in this study and some necessary pre-processing are first detailed. Diagnos-

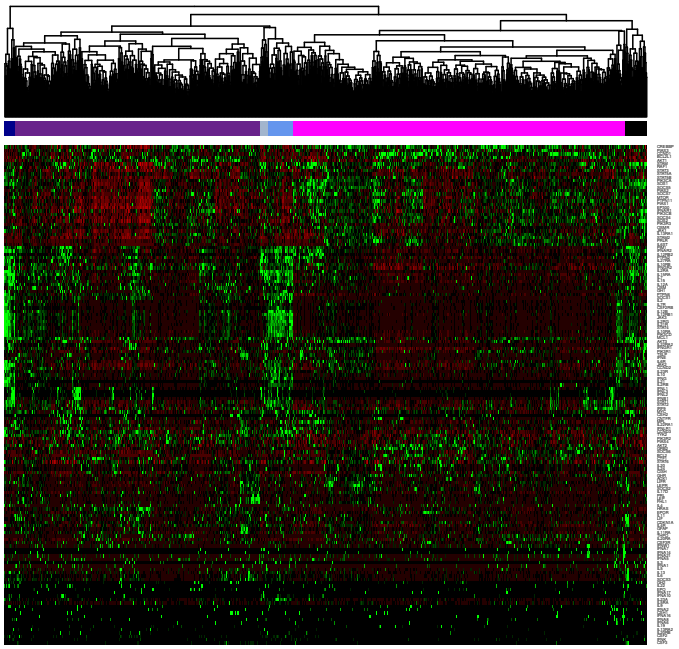
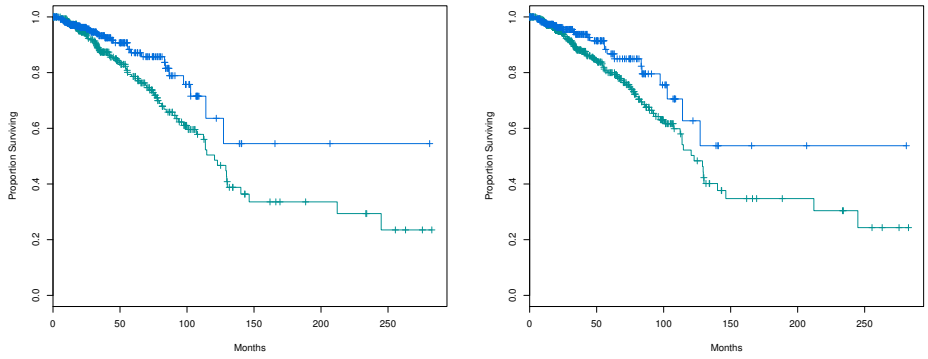


Figure 4.2: Heatmap of genes involved in the Jak-STAT signaling pathway in KEGG.

tic WSI images, clinical data, and gene expression from 710 patients from TCGA-BRCA were used in this study. RNA-seq (v2) median-centered z-scores of gene expression for over 16,000 genes for each patient were retrieved from TCGA using cBioPortal's R package "cgdsr". The z-scores were thresholded to be within -4 to 4 to limit the bias of large outliers. For each patient, up to 15 1000 $\times$ 1000 pixel patches from each WSI were manually selected for analysis, avoiding possible contamination by artifacts of blurred regions or tissue folds. For patients with multiple diagnostic WSIs, only the first slide was used. The WSIs for all patients included in the study were imaged at 40$\times$ magnification.

4.3 Unsupervised Image-Based Clustering Analysis

The first experiment for associating image data with genomics uses unsupervised clustering to group patients from the TCGA-BRCA dataset into data-driven clusters and then to search for genes associated with these clus-



(a) All features (Log-rank test: $p = 0.00975$)

(b) Texture (Log-rank test: $p = 0.0316$)

(c) Area and Shape (Log-rank test: $p = 0.905$)

(d) Intensity (Log-rank test: $p = 0.0397$)

Figure 4.3: Survival curves for the two separated image-based clusters for different feature subsets.

ters based on gene expression and mutation data. The proposed H&E analysis pipeline was applied to the TCGA-BRCA dataset of 710 patients to transform the WSI of each patient into a single feature vector capturing the distribution of nuclear and cellular features in the WSI.

The summarized feature vectors of all patients were separated into clusters using a Gaussian mixture model. To determine if the clusters were meaningful, they were tested for significant association with outlook among the patients of each group. To remove the bias of outliers, which may represent images of significantly poor quality, they were removed prior to clustering using an isolation forest classifier and added again for subsequent analysis. To gain insights into the significance of different features, patients were sep-

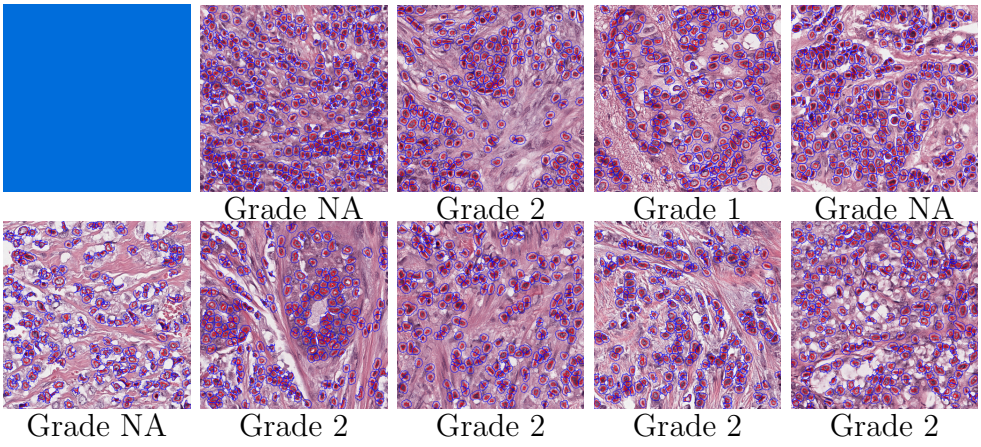
Table 4.1: KEGG pathways associated with genes that were differentially expressed across image-based clusters.

Pathway abbrv.	Pathway name	p-value	FDR
PImm	Primary Immunodeficiency	$4.4 * 10^{-8}$	$5.4 * 10^{-5}$
Tcell	T cell receptor signaling pathway	$6.5 * 10^{-7}$	0.0008
NKcell	Natural killer cell mediated cytotoxicity	$1.4 * 10^{-4}$	0.17
Cyto	Cytokine-cytokine receptor interaction	$3.0 * 10^{-4}$	0.37
Bcell	B cell receptor signaling pathway	$7.9 * 10^{-4}$	0.98
CAM	Cell adhesion molecules	0.0076	9.1
NFK	NF-kappa B signaling pathway	0.013	14.7
Phag	Fc gamma R-mediated phagocytosis	0.045	44
Jak	Jak-STAT signaling pathway	0.083	66.05

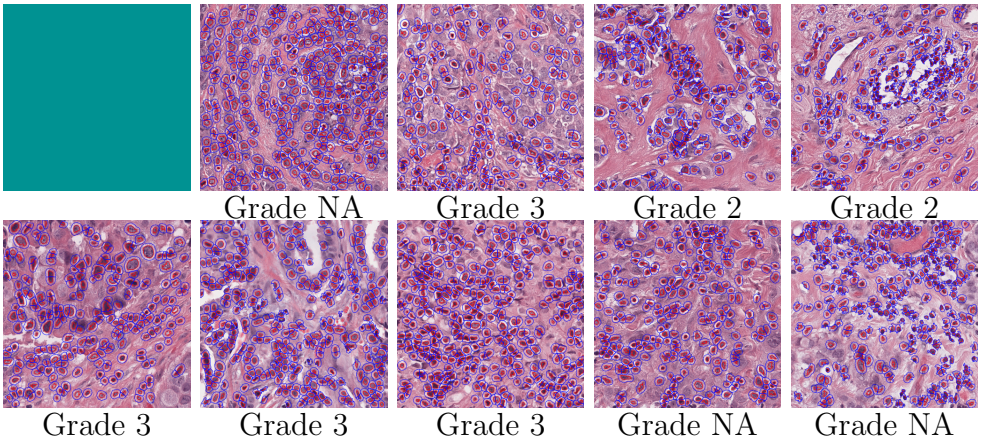
arated into two clusters based on different subgroups of image features. The resulting survival curves for clusters based upon texture features, area and shape features, intensity features, and all features are shown in Fig. 4.3. Texture features, which describe the homogeneity of the nucleus and surrounding cytoplasm, showed a strong association with survival. The intuition of this finding is confirmed by the routine Nottingham Histologic Score, which incorporates nuclear pleomorphism, such as the prominence of the nucleoli and presence of vesicles, as an important prognostic indicator. Intensity features, which capture the distribution of hematoxylin and eosin intensities within the nuclei and cells and are also quantitative measures of nuclear pleomorphism, also showed a significant association with survival.

Clustering with all features revealed two clusters with the most distinct survival curves. Most of the features distinguishing these two clusters were indeed texture and intensity features, primarily of the eosin channel. The clusters derived from area and shape features, while also being metrics for nuclear pleomorphism, showed little association with survival, which may be due to a still greater need for accurate and robust segmentation, to which these features are more sensitive. Fig. 4.4 shows patches from the WSIs of patients whose image feature vector was closest to each cluster center.

We used the R package SAM [53] to test for significant associations between our image-based clusters and the expression levels of 16,000 genes. The analysis reported 254 genes that were significantly differentially expressed (with a false discovery rate [FDR] less than 0.05) between the two clusters for which all features were used. To visualize the pattern of expression of these



(a) Image-Based Cluster with Better Prognosis



(b) Image-Based Cluster with Poor Prognosis

Figure 4.4: Representative patches from nine different patients for the two clusters derived using all features. Patches for the poor prognosis group are shown in the bottom two rows, along with the green color patch to match with other figures, and patches for the better prognosis group are shown on the top two rows, along with a blue color patch. The nuclear grades, if provided in the pathologist's report, for the patient corresponding to each are given below the patch.

genes across the two clusters, the median-centered z-scores were grouped using hierarchical clustering within each image-based cluster and are shown in Fig. 4.5. In the heatmap, each row corresponds to one of the 254 differentially expressed genes, and each column corresponds to a patient. The patients are grouped by image cluster, indicated by the green and blue bars across the top of the heatmap, and then clustered hierarchically within each image cluster to help visualize similar expression patterns. The genes are clustered hierarchically to help with visualization as well. A strong pattern of over-

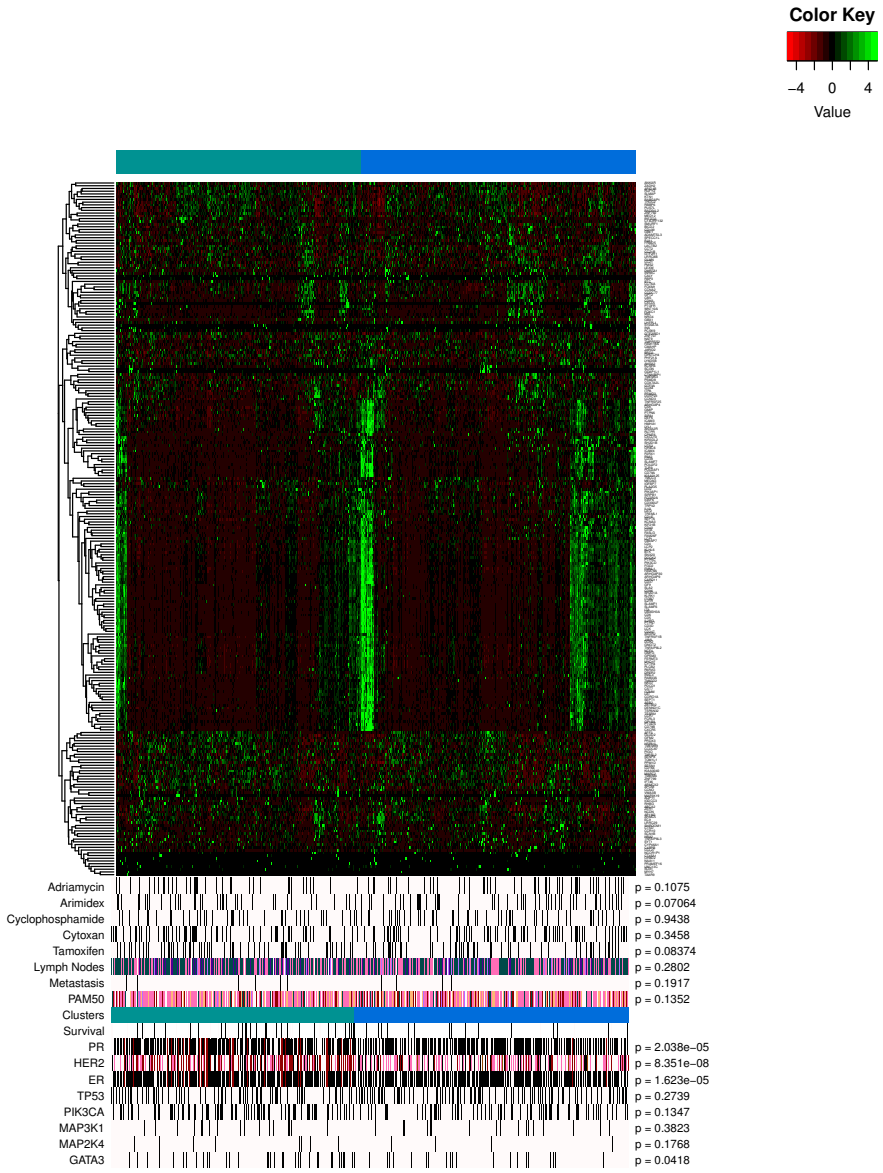


Figure 4.5: Gene expression of significantly differentially expressed genes (FDR ≤ 0.05) between the two image-based clusters using all image features. The cluster with better survival (blue) shows two groups of patients with significantly higher gene expression. It is also evident that a large group of the selected genes are highly correlated for a given patient.

expression for nearly half of these genes can be seen for the group of better prognoses.

The table below the heatmap in Fig. 4.5 shows the association of these clusters with clinical markers, the administration of several drugs, and several gene mutations. Each column in the table is aligned to the same column in

Table 4.2: List of several significantly differentially expressed genes (FDR ≤ 0.05) between the discovered image-based clusters and significantly associated KEGG pathway membership.

Gene ID	PImm	Tcell	NKcell	Cyto	Bcell	CAM	NFK	Phag	Jak	Contrast-1	Contrast-2
CIITA										2.059	-2.453
PTPRC										1.801	-2.146
CD3D										1.982	-2.361
LCK										2.278	-2.714
IL2RG										2.138	-2.547
CD4										1.803	-2.148
CD79A										2.021	-2.408
CD40										2.277	-2.713
BTK										2.019	-2.405
LAT										2.071	-2.467
CARD11										2.099	-2.501
PIK3CD										1.963	-2.339
CD247										2.062	-2.456
IL10										1.885	-2.246
PDCD1										2.052	-2.445
LCP2										1.873	-2.232
PTPN6										1.939	-2.31
SH2D1A										1.894	-2.256
KLRK1										1.802	-2.147
FASLG										1.959	-2.334
SH2D1B										1.794	-2.137
TNFRSF25										2.11	-2.513
IL21R										1.951	-2.324
CCL27										1.88	-2.239
CXCL10										1.881	-2.241
CCR7										1.8	-2.144
TNFRSF1B										1.81	-2.157
CXCR5										1.879	-2.238
CCR10										2.065	-2.46
LTA										2.323	-2.768
CARD11										2.099	-2.501
PTPN6										1.939	-2.31
PIK3AP1										1.835	-2.186
CD79B										2.059	-2.453
ITGB7										2.448	-2.492
ICAM3										2.218	-2.643
CD6										2.091	-2.492
CDH3										1.885	-2.245
DOCK2										1.823	-2.172
FCGR2A										2	-2.383
CCND3										1.997	-2.379

the heatmap and corresponds to the same patient. The status of the primary clinical immunohistochemistry markers used in breast pathology (estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor-2 (HER2)) are displayed, with black indicating positive status and white indicating negative status. These markers show a strong association with the two clusters, which implies that the difference in survival of the two clusters is in some way related to known cancer biomarkers. Mutations are also displayed, with black indicating the presence of a mutation. Of the five mutations associated with breast cancer that were considered in the analysis, only GATA3 showed a significant association ($p = 0.0418$). The five drugs most commonly administered to breast cancer patients were tested for association, and no significant associations were found, implying that drug

administration is likely not the cause of the survival differences of the two groups. In the table, black indicates a drug was administered and white indicates it was not. Neither lymph node status, PAM50 subtype, nor the presence of metastasis showed a strong association with the clusters.

To gain insight into the function of these differentially expressed genes, we used DAVID [60] to search for KEGG pathways in which significant subsets of these genes are present. The top five most strongly associated pathways were: primary immunodeficiency, T-cell receptor signaling pathway, natural killer cell mediated cytotoxicity, cytokine-cytokine receptor interaction, and B-cell receptor signaling pathway. Significance values of association and used abbreviations for each pathway are given in Table 4.1. Each of these pathways was associated with a p-value less than 0.05, while primary immunodeficiency and T-cell receptor signaling pathway were also associated with a FDR less than 0.05. We incrementally raised the admissible FDR for our gene set, up to 3.475, and continued to observe these pathways associated at similar levels of significance. Several other pathways implicated in cancer were also discovered, but these pathways had a significantly higher FDR.

The genes involved in each of these pathways that were significantly differentially expressed across the two clusters are shown in Table 4.2. Each column in the table corresponds to a pathway listed in Table 4.1. The presence of a colored square at a particular row and column indicates that the gene of its row is involved in the pathway of its column. The contrast values, given in the last two columns, are the standardized mean difference between the expression levels of the two clusters. Contrast-1 corresponds to difference for the improved prognosis group and contrast-2 corresponds to the difference for the poor prognosis group. The positive contrast values (contrast-1) for the improved prognosis group compared to the negative contrast values (contrast-2) for the poor prognosis group correspond with the higher levels of expression of the selected genes in Fig. 4.5.

This analysis provided possible insights for further exploration of the influence of these pathways on nuclear pleomorphism, as measured by various texture and intensity features. The most significantly associated pathways discovered are related to the immune system, which is crucial for recognizing and killing cancer cells. Other pathways also have a role in cancer: the CAM pathway can affect metastasis and both the NF-kappa B signaling and the Jak-STAT signaling pathways affect cell growth.

4.4 Supervised Image-Based Clustering with Glmnet

A concern of the previous experiment is how well the clusters will generalize to new data. Since unsupervised learning is not being driven by a class label, such as patient outlook, each feature in the data is given equal weight in the clustering assignment and the algorithm is unable to differentiate between features that are informative of outlook.

Supervised learning algorithms use a desired class assignment, or response variable, to learn which features in the data are discriminative and what parameters in the assumed model are optimal for discriminating the response variable. In the case of right-censored data like survival data, instead of a class label for each patient, each patient has a time-of-last-follow-up y and an event indicator of status (dead or alive) δ, denoted (x_i, y_i, δ_i) for patient i. There are a variety of models for fitting right-censored survival data, but a commonly used model is the Cox proportional hazard. This model assumes that the hazard for each patient follows a baseline curve $h_0(t)$ scaled by an exponential of the linear regressor:

$$h_i(t) = h_0(t)e^{x_i^T \beta}, \tag{4.3}$$

where β is the vector of linear parameters of the model to be learned. These parameters are chosen by maximizing the likelihood:

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_{j(i)}^T \beta}}{\sum_{j \in R_i} e^{x_j^T \beta}}. \tag{4.4}$$

A tool for performing this optimization is Glmnet [61] (notation is adapted from this reference), which minimizes the negative log-likelihood over the data under a regularization penalty on the l_1 and l_2 norms, a mixture of the LASSO and ridge penalties, called the elastic-net penalty.

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right]. \tag{4.5}$$

This penalty trades off between the number of features allowed in the model, enforced by the LASSO penalty, and the squared magnitude of the selected features, the ridge penalty via α.

Glmnet was applied to the dataset of TCGA-BRCA patients to learn a predictive model for separating patients into poor and better prognoses groups. The patients were first split randomly into two groups: 80% were used for training the model, and 20% were used for testing the model. The training data were shuffled 15 times, and for each shuffle, 75% were used to train the model and 25% were used to validate it. The penalty term λ was chosen for each shuffle using 10-fold cross validation on the subset of 75% of the training data. The value $\alpha = 0.8$ was chosen manually, as it tended to produce the best performance on validation sets. The model with the best separation of survival curves, according to the p-value, was chosen.

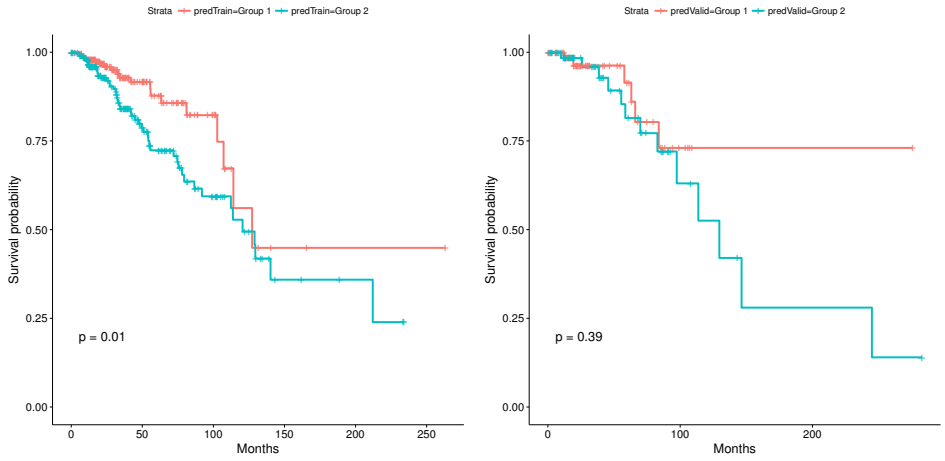
For each set of training data, Glmnet trained a hazards model for the data, and a risk index for each patient in the training data was computed. Following another similar approach for non-small cell lung cancer prognostic prediction [28], the median risk index of the data was chosen as the separating threshold for the two clusters.

Models were generated for both data that had been stain-normalized and data that had not. The survival curves of the discriminated groups without stain-normalization are shown in Fig. 4.6. The survival curves of the discriminated groups with stain-normalization are shown in Fig. 4.7. Without stain normalization, Glmnet was able to learn a model that separated the training data, but it could not generalize well to the validation or test sets. It is possible that this model is fitting too closely to the characteristics of the data acquisition process of particular TSSs, such as stain concentration and sample thickness. Although stain normalization can introduce its own artifacts, in this scenario it is able to help the model learn more robust features. The validation and test sets show much better discrimination, though the discrimination still does not generalize completely to validation and testing sets.

An advantage of Glmnet, with its elastic-net penalty, is that it produces a sparse subset of nonzero coefficients in β via the l_1 penalty, which is informative of the importance of each feature. The selected features, along with their corresponding coefficients in β in the regression model, for stain-normalized images are given in Table 4.3. Since the features that are passed to Glmnet are normalized to zero mean and unit variance, the magnitude of the coefficients can be thought of as a measure of importance in the model. Notably, most features with a large coefficient are Zernike moments of the

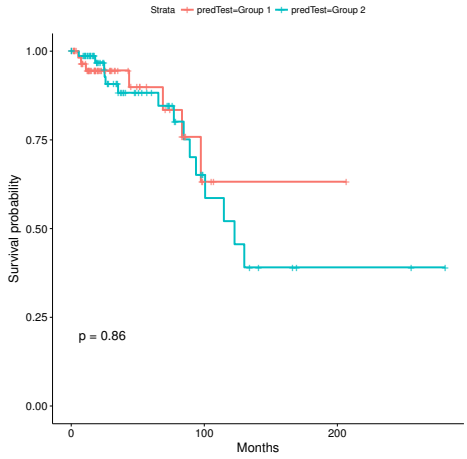
Table 4.3: Image features selected by Glmnet to separate prognoses groups.

Image Feature	Coefficient	Image Feature	Coefficient
AreaShape_Zernike_9_9	−5.4289	P40_AreaShape_Zernike_9_9	−9.5191
Cells_AreaShape_Eccentricity	0.0586	P40_Cells_AreaShape_Eccentricity	0.4422
MaskedCells_Intensity_MassDisplacement_e_img_invert	−0.0752	P40_Cells_AreaShape_Zernike_2_2	0.9638
Std_AreaShape_MajorAxisLength	0.0027	P40_MaskedCells_Intensity_LowerQuartileIntensity_e_img_invert	0.1112
Std_AreaShape_Zernike_7_1	20.6470	P40_MaskedCells_Intensity_MeanIntensity_e_img_invert	0.1176
Std_AreaShape_Zernike_7_3	−64.4858	P40_MaskedCells_Intensity_MedianIntensity_e_img_invert	0.1131
Std_AreaShape_Zernike_8_6	−105.5879	P50_AreaShape_MedianRadius	0.0890
Std_Cells_AreaShape_Extent	0.5160	P50_AreaShape_Zernike_6_2	22.9912
Std_Cells_AreaShape_MajorAxisLength	0.0269	P50_AreaShape_Zernike_9_9	−35.7028
Std_Cells_AreaShape_MaxFeretDiameter	0.0188	P50_Cells_AreaShape_Eccentricity	0.3506
Std_Cells_AreaShape_MinFeretDiameter	0.0104	P50_Cells_AreaShape_Zernike_9_3	9.0434
Std_Cells_AreaShape_MinorAxisLength	0.0069	P50_MaskedCells_Intensity_MassDisplacement_e_img_invert	−0.0047
Std_Cells_AreaShape_Zernike_3_1	46.4751	P60_AreaShape_Zernike_6_0	−25.3143
Std_Cells_AreaShape_Zernike_7_3	43.1110	P60_AreaShape_Zernike_9_9	−88.3824
Std_Cells_AreaShape_Zernike_8_4	60.7172	P60_Cells_AreaShape_Eccentricity	0.6175
Std_MaskedCells_Intensity_IntegratedIntensityEdge_e_img_invert	0.0040	P60_MaskedCells_Intensity_MassDisplacement_e_img_invert	−0.0279
Std_MaskedCells_Intensity_IntegratedIntensity_e_img_invert	0.0005	P70_AreaShape_Zernike_5_1	12.9438
Std_MaskedCells_Intensity_MassDisplacement_e_img_invert	−0.1434	P70_AreaShape_Zernike_9_9	−9.5144
Std_MaskedCells_Texture_Entropy_e_img_invert_3_avg	0.0224	P70_Cells_AreaShape_Eccentricity	0.6339
P10_AreaShape_Zernike_8_0	−174.8484	P70_Cells_AreaShape_Zernike_8_4	9.6260
P10_Texture_Entropy_e_img_invert_3_avg	0.0883	P70_MaskedCells_Intensity_MassDisplacement_e_img_invert	−0.0668
P10_MaskedCells_Intensity_LowerQuartileIntensity_e_img_invert	0.1847	P80_AreaShape_Zernike_5_1	13.5263
P10_MaskedCells_Intensity_MeanIntensity_e_img_invert	0.2590	P80_AreaShape_Zernike_7_1	9.5121
P10_MaskedCells_Intensity_MedianIntensity_e_img_invert	0.1484	P80_AreaShape_Zernike_9_9	−4.2373
P20_MaskedCells_Intensity_LowerQuartileIntensity_e_img_invert	0.2261	P80_Cells_AreaShape_Eccentricity	0.4338
P20_MaskedCells_Intensity_MeanIntensity_e_img_invert	0.2249	P80_MaskedCells_Intensity_MassDisplacement_e_img_invert	−0.0722
P20_MaskedCells_Intensity_MedianIntensity_e_img_invert	0.1051	P90_AreaShape_MedianRadius	0.0080
P30_AreaShape_Zernike_8_0	−25.7123	P90_AreaShape_Zernike_8_6	−46.7452
P30_MaskedCells_Intensity_LowerQuartileIntensity_e_img_invert	0.1698	P90_AreaShape_Zernike_9_9	−2.4967
P30_MaskedCells_Intensity_MeanIntensity_e_img_invert	0.2304	P90_Cells_AreaShape_Zernike_3_1	6.7267
P30_MaskedCells_Intensity_MedianIntensity_e_img_invert	0.0790	P90_Cells_AreaShape_Zernike_8_4	46.1625
P40_AreaShape_Zernike_8_0	−30.5062	P90_MaskedCells_Intensity_MassDisplacement_e_img_invert	−0.0681

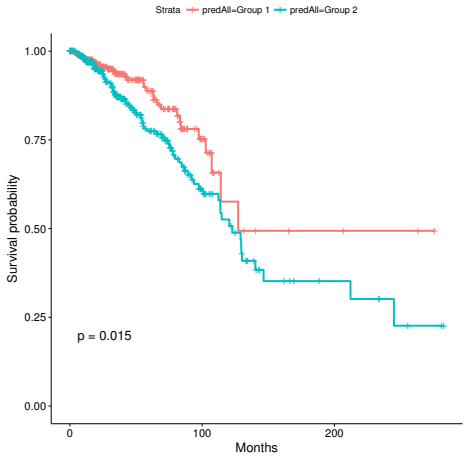


(a) Training Results

(b) Validation Results



(c) Testing Results



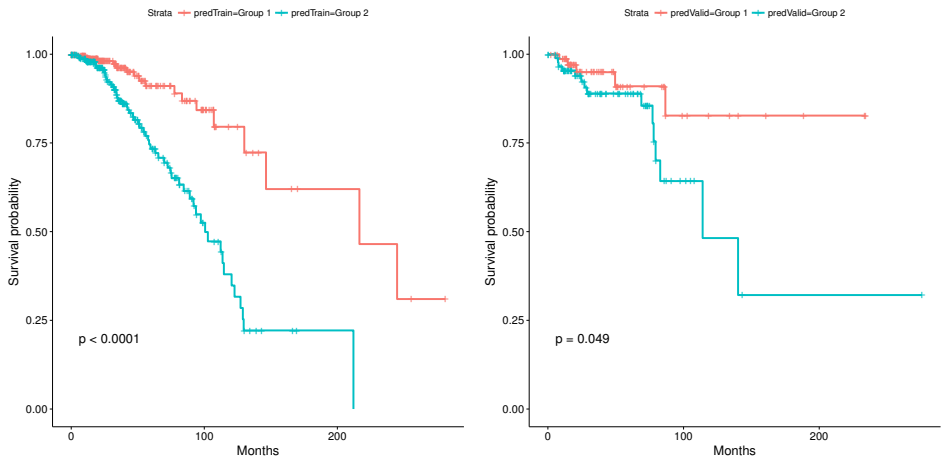
(d) All Results

Figure 4.6: Survival curves for learned hazards model using Glmnet on TCGA-BRCA images that were not stain-normalized.

cells and nuclei, which are invariant descriptors of shape. In contrast to the unsupervised GMM clusters, shape features are discovered to be the most informative of survival by Glmnet. Other features of texture and intensity are also present in the selected subset, but are significantly less informative.

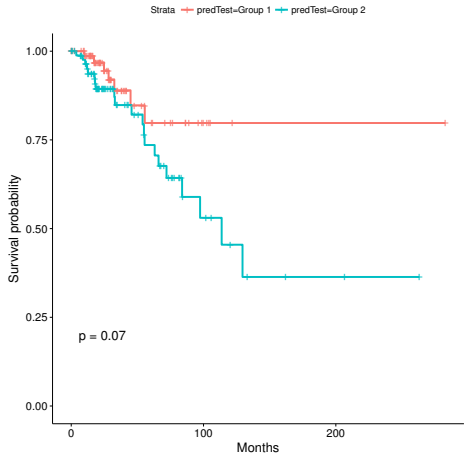
4.4.1 Luminal A Subtype

The same experiment was performed the subset of TCGA-BRCA patients that belonged to the luminal A molecular subtype. This subtype tends to be ER-positive, HER2-negative, and of a low tumor grade, and tends to have

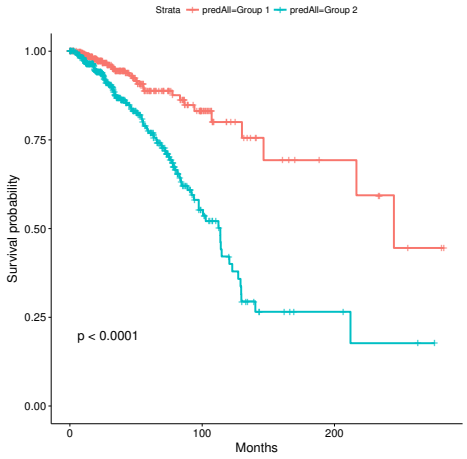


(a) Training Results

(b) Validation Results



(c) Testing Results



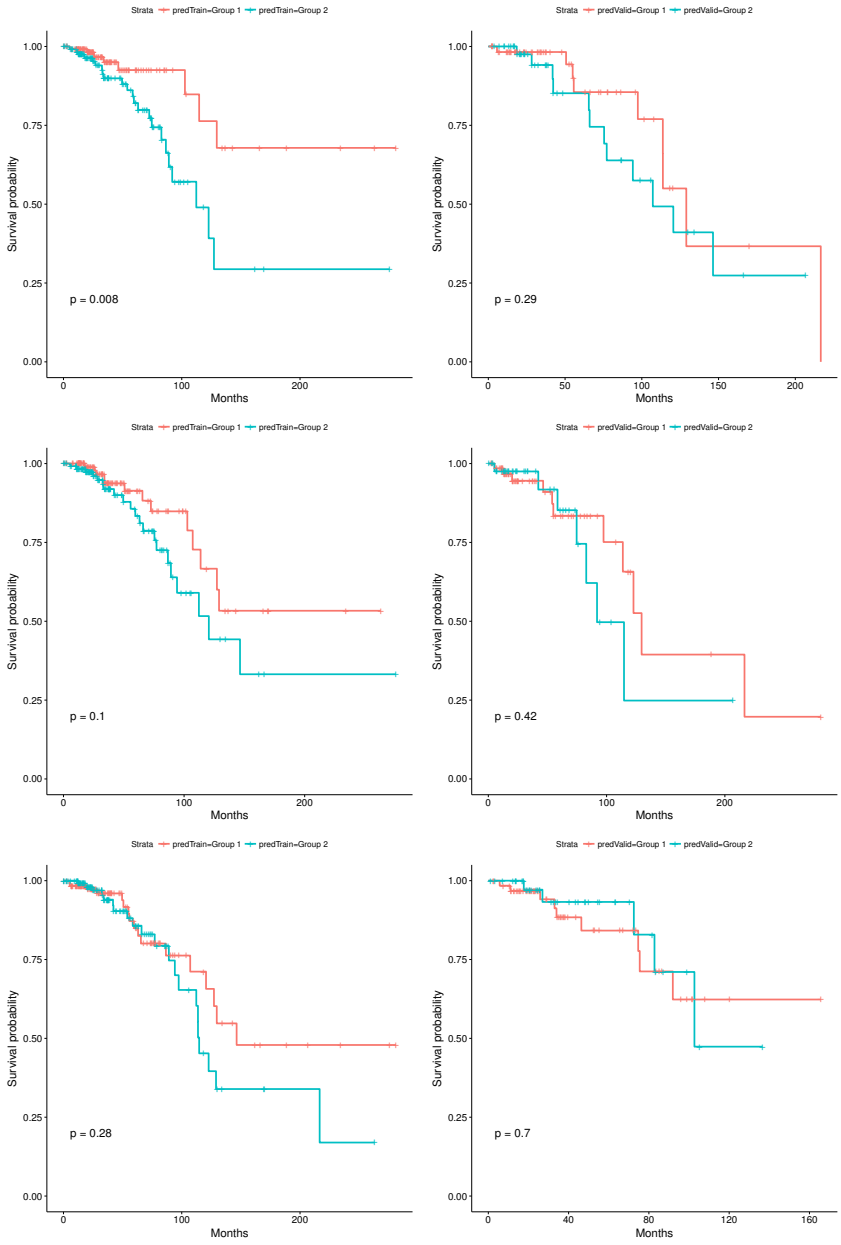
(d) All Results

Figure 4.7: Survival curves of poor and better prognoses groups according to the learned hazards model using Glmnet on stain-normalized TCGA-BRCA images.

the best prognosis of molecular subtypes. This subtype was chosen among the five primarily because it constituted the largest subset of patients, with 367 members.

Since restricting the patient set to this subtype drastically reduced the number of patients for which to learn a model, to evaluate the model, the dataset was split randomly 15 times between 70% training and 30% testing. Results of three of the 15 iterations are shown in Fig. 4.8. Similar to the experiments using all patients, the model is generally able to separate the training data well, but fails to generalize to held-out testing data. These results indicate that more informative features are still needed to capture

any differences in prognoses within the luminal A subtype. The same is likely true for the other subtypes as well.



(a) Training Results (b) Testing Results

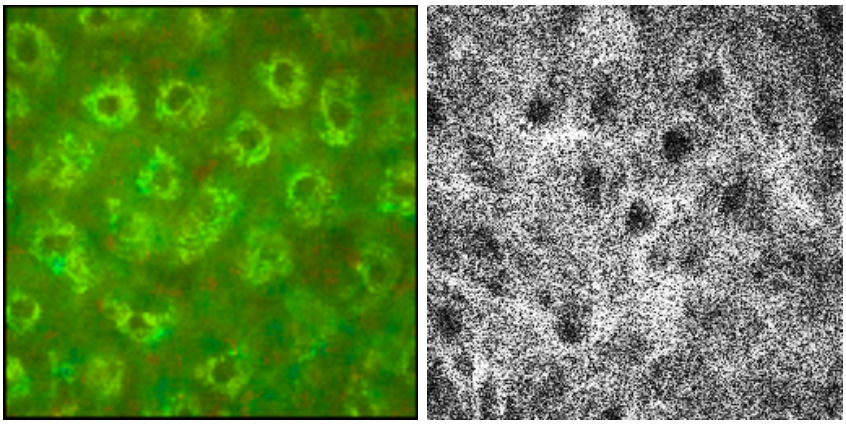
Figure 4.8: Survival curves of good and poor prognoses groups according to the learned hazards model using Glmnet on stain-normalized TCGA-BRCA images of the luminal A subtype. Each row shows the resulting curves of a different random split of the data into (a) training and (b) testing sets. The inability of the model to generalize well to testing data for each different partitioning of the dataset indicates the need for the development of more informative image features.

CHAPTER 5

MULTIMODAL IMAGE ANALYSIS

Though hematoxylin and eosin staining of tissue slides remains the gold standard for diagnosis for most cancers, it is limited in what information it can provide about the cellular processes inside a tumor. Other modalities, such as fluorescence lifetime imaging (FLIM), exist that can reveal unique quantitative information about these processes. FLIM can provided quantitative measurements that are related to the metabolic activity within a sample, and it has been shown to be discriminative of processes of cell death, namely apoptosis and necrosis. Apoptosis, a natural response of cells to nonviable conditions, requires an increase in metabolic activity, whereas necrosis, the uncontrolled and inflammatory death of cells, generally does not. The type of death that cells within a tumor experience is informative of the stage of cancer, since suppression of apoptosis is a key hallmark [1]. Furthermore, FLIM could be useful for clinical applications, such as testing the efficacy of drugs and treatment in inducing apoptosis. Optical coherence microscopy can complement FLIM by providing structural information about the environment. Additionally, these modalities can be used to image live cells, allowing for the observation of progression of cancer over time, and they can even be imaged *in vivo*. Recently, multimodal imaging systems have been proposed and developed, making possible quantitative analysis of a tumor via multiple imaging modalities that are spatially and temporally registered. In particular, a system has been developed that can image FLIM, OCM, and multiphoton microscopy (MPM) simultaneously *in vivo* [62], and this system is the focus of this thesis.

Processing the acquired images by human observation alone is not feasible and would likely not fully utilize the quantitative aspect of these several modalities. Therefore, it is desirable to develop image processing algorithms to extract meaningful information from these images and machine learning algorithms to learn how best to leverage the unique information provided



(a) FLIM (b) OCM

Figure 5.1: Multimodal images from necrosis-induced cells after 3 hours at a depth of 0 μm. Histogram equalization has been applied to the OCM image for visualization.

by each modality for specific tasks. In particular, this thesis considers the problem of detecting and differentiating apoptosis and necrosis and proposes an algorithm for detecting and segmenting cells, as well as a statistical model for the detection problem.

5.1 Detection of Apoptosis and Necrosis in Cancerous Cells Using MPM-OCM-FLIM Multimodal Data

Images of living engineered skin cells were collected by the imaging system *in vivo* over 24 hours for three studies: apoptosis, necrosis, and a control study of homeostasis. The modalities of the imager were OCM, FLIM photon count, and FLIM curve-fit parameter. An example multimodal image from a cancer cell study is shown in Fig. 5.1. The first stage of the detection pipeline is the tracking of cells and the segmentation of cell cytoplasm. Once the regions of cell cytoplasm are identified, the pixel values within these regions can be used to detect apoptosis, necrosis, or homeostasis, since the cytoplasm is where most of the metabolic activity in the cell occurs, as opposed to the nucleus.

5.1.1 Cell Tracking and Segmentation

Delineating clear cell boundaries from the multimodal data poses a challenge due to noise. However, the consistency of images over time can be used advantageously to combat this challenge. We registered FLIM photon count images at the same depth across different time stamps using a mutual information cost function [63] and a (1+1) evolutionary optimizing strategy [64], assuming a rigid transform with the first image of the time-lapse. With accurate registration, an average of the time-lapse images can be computed to aid in segmentation by increasing the signal-to-noise ratio and the contrast. This segmentation algorithm also requires initial seeds of cytoplasmic regions, which are currently labeled manually, though an algorithm to robustly and automatically detect nuclei to be used could be added. The segmentation algorithm morphs the region initialized by the seed using active contours [65]. A segmentation mask is generated as the final output, which can be refined using morphological operations and size constraints. The mask is then transformed to the image at each different time stamp using the previously derived registration mapping. Example outputs for apoptosis (top), necrosis (bottom), and control (middle), or homeostasis, studies are shown in Fig. 5.2. As can be seen, the current tracking and segmentation method does reasonably well, but struggles, especially at the 24 hour mark, since much has changed in the environment since the last time stamp and the image content becomes diffuse.

5.1.2 GMM-Based Classification

Once the cytoplasmic pixels are segmented, trends in the three modalities that separate the two processes of cell death and the control state of homeostasis can be investigated. Clearly distinguishable trajectories through the three dimensional data space over time can be seen, as shown in Fig. 5.3, when all pixels started at roughly the same location in the space. In images from the apoptosis study, a significant increase in the FLIM curve parameter is seen over time, which is consistent with other studies that showed increased metabolic activity during apoptosis. By the end of the observation of 24 hours, the data of the three studies had separated into well-delineated clusters. The FLIM photon count and OCM values can also be seen to play

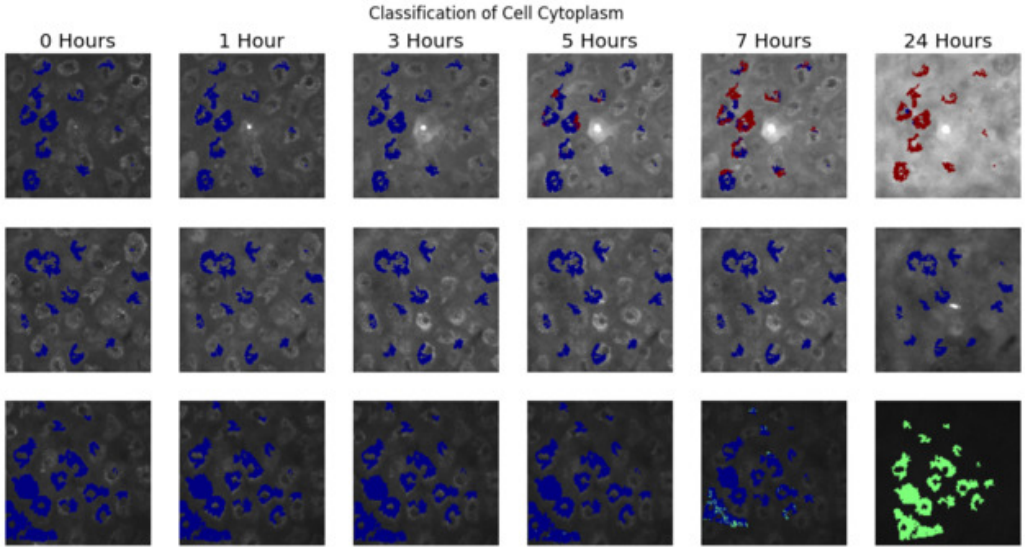


Figure 5.2: Evolution of cell tracking and cytoplasm segmentation over time for apoptosis (top row), homeostasis (middle row), and necrosis (bottom row) studies. The colors indicate the predicted labels of each segmented pixel by the GMM at each time stamp (red = apoptosis, blue = control, green = necrosis).

a role in separation of the three studies.

A naive approach that demonstrates the separability of these three studies was applied using a GMM. Each data point was assumed to be generated from a Gaussian distribution of its corresponding study, or class. The model was trained using data from the last observation at 24 hours. The parameters of the trained model were then used to classify the data points at each time stamp. Fig. 5.3 shows that some pixels from the apoptosis study were assigned correctly even at the 7 hour time stamp.

Further studies need to be conducted to ensure that the variation between processes is not a result of variation in the image acquisition process. Variation such as image source intensity can significantly affect the location of the clusters in the data space and their trajectory, which would compromise the effectiveness of this clustering method.

5.1.3 Mathematical Formulation of Detection Problem

Although the GMM model is effective at classifying the three studies in this example, it is likely that this model would not be robust against variation

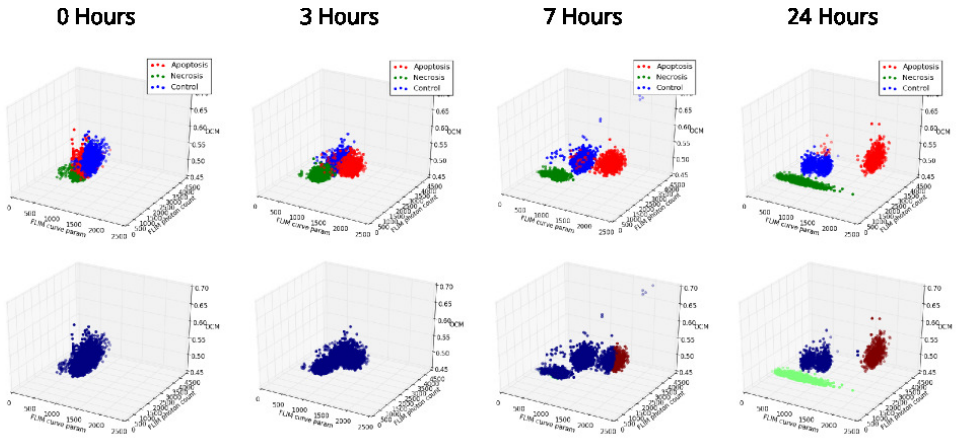


Figure 5.3: Scatter plot of the evolution of cytoplasmic pixels from necrosis (green), apoptosis (red), and control (blue) studies over time. As time progresses, a clear distinction can be seen between the three clusters. The upper row shows the true labels of the three studies. The lower row shows the estimated labels using the GMM model.

across experiments, in particular variation in the image acquisition process. Additionally, as can be seen in the results of Fig. 5.3, separation of the studies can be identified earlier than the 24 hour time stamp, which the GMM model does not fully leverage. Furthermore, in actual *in vivo* experiments, the time stamps will not be known, and so the stage that a cell is in within a process must be estimated in addition to the process itself. The subject of future work will be to formulate a more precise mathematical model that accounts for these possible sources of variation, handles the estimation of the stage in the particular process, and quantifies precisely the uncertainty of detection.

Consider a simple model of two possible trajectories of cells through the three-dimensional data space, where one trajectory is a form of cell death, either necrosis or apoptosis, and one is homeostasis. Let the random variable $X_t \in \mathbb{R}^3$ be the position in the space at time stamp t of the average of all pixels in the cell under consideration. Assume that, if cell death was induced, it began at $t = 0$. Let the random vector $\vec{X} = \{X_0, X_1, \ldots, X_{N-1}\}$ be the vector of data points of each time stamp, representing the trajectory of the cell over time, and N be the number of time stamps collected. Let the random variable C be the label for the cell process ($C = 0$ represents homeostasis and $C = 1$ represents cell death). This problem is essentially a hypothesis detection problem, for which the maximum *a posteriori* (MAP) estimate $\hat{C}$

of C, which is optimal for the 0-1 loss, is given by

$$\hat{C}_{\text{MAP}} = \arg\max_c P(C = c|\vec{X}). \tag{5.1}$$

If the prior probability for each label is assumed to be equal, then the MAP estimate is the maximum likelihood estimate (MLE) via Bayes' rule:

$$\hat{C}_{\text{MLE}} = \arg\max_c P(\vec{X}|C = c). \tag{5.2}$$

If the expected trajectory for homeostasis remains close to some mean point, and the trajectory for cell death follows some known trajectory, the two observed trajectories can be modeled as following one of these two known trajectories in the presence of independent Gaussian noise (assuming the mean of homeostasis has been removed):

- Homeostasis: $X_t \sim \mathcal{N}(0, \sigma_0)$
- Cell death: $X_t \sim \mathcal{N}(\mu_t, \sigma_1)$

Given this model, since the noise at each time stamp is independent, the likelihood probability for each process is simply the product of the likelihoods of each time stamp:

$$P(\vec{X}|C) = \prod_{t=0}^{N-1} P(X_t|C). \tag{5.3}$$

Given the stated models for the two possible trajectories, the MLE is a simple log ratio test,

$$\hat{C}_{\text{MLE}} = \frac{P(\vec{X}|C = 1)}{P(\vec{X}|C = 0)} \underset{0}{\overset{1}{\gtrless}} 1, \tag{5.4}$$

$$= \sum_{t=0}^{N-1} \left(\frac{X_t^2}{\sigma_0^2} - \frac{(X_t - \mu_t)^2}{\sigma_1^2} \right) \underset{0}{\overset{1}{\lessgtr}} \sum_{t=0}^{N-1} \left(\log \sqrt{2\pi}\sigma_1 - \log \sqrt{2\pi}\sigma_0 \right). \tag{5.5}$$

In a real *in vivo* scenario, the time point t of an image would not be known, since there would be no start time to reference. To account for this, the model must be enhanced by the introduction of a random variable, T, to represent the stage, or time stamp, of the process of cell death at the first observation. Imposing a uniform prior on $T_0 \in \{0, 1, \ldots, N - \tau - 1\}$, thereby assuming that the cell is equally likely to be in any stage of cell death at the

first observation, assuming it is experiencing cell death, and assuming that τ time stamps are observed and N is the last modeled stage in the process, the new posterior probability of the label can be written as

$$P(C|\vec{X}) = \sum_{t_0=0}^{N-\tau-1} P(C, T_0 = t_0|\vec{X}) \propto \sum_{t_0=0}^{N-\tau-1} P(\vec{X}|C, T_0 = t_0), \tag{5.6}$$

$$= \sum_{t_0=0}^{N-\tau-1} \prod_{t=t_0}^{t_0+\tau} P(X_t|C, T_0 = t_0). \tag{5.7}$$

From here, the model can be further enhanced to account for a third trajectory for cell death, distinguishing between apoptosis and necrosis, and other prior knowledge. Other models for the trajectories may be better suited, such as a Markov chain, which could replace the simple model of a known trajectory with Gaussian noise. Future work will revise and enhance this model, from which the optimal decision rule and associated confidences can be derived, and test the model on controlled *in vivo* data and real *ex vivo* data.

CHAPTER 6

CONCLUSION

There is still much to be understood about the connection between genotype and phenotype in cancer, and more sophisticated tools for image and genomic analysis will be required. This thesis has proposed a computational pipeline for analyzing H&E images by extracting features and locations of cells and their nuclei that could serve as a building block for the development of such tools, such as graphical models that capture the spatial layout of the tumor. Additionally, this thesis has proposed a CNN to perform robust nuclear segmentation, a key component of histopathological image analysis, and has demonstrated its effectiveness across a variety of WSIs compared to other commonly used tools. To demonstrate the efficacy of this pipeline, it was applied to 710 TCGA-BRCA patients to extract nuclear and cellular features. Clustering patients into subgroups evidenced that the computed features have a meaningful relationship with patient outlook. The lack of any necessary parameter tuning in the segmentation step and the ease of use of this pipeline should help researchers working with H&E images to perform more reliable analysis, whether WSIs or biopsies, without needing to understand image processing techniques. This pipeline can also be easily parallelized on a cluster to run on separate partitions of large datasets. Additionally, continuing to increase the training set of nuclei patches will help to improve the segmentation accuracy of the proposed CNN.

This thesis has also demonstrated a means of analyzing connections between computed image features and genomic data, specifically gene expression, by searching for significantly differentially expressed genes across image-based clusters using SAM and then using DAVID to find associated KEGG pathways. This analysis provided possible insights for further exploration of the influence of these pathways on nuclear pleomorphism, as measured by various texture and intensity features. Though we have demonstrated the analysis of our H&E pipeline with a particular method for associating im-

age features with genomics, the use of the pipeline is not restricted to this procedure, and other means of investigating connections could be explored. We envision this pipeline being applied to H&E datasets for a variety of subsequent imaging-genomic analysis.

Since cells are living, dynamic entities that experience various cycles and processes over time, and biomarkers such as gene expression change through these processes, other modalities that can image live cells will be necessary for future research. This thesis has demonstrated how image processing can aid multimodal imaging systems in obtaining quantitative measurements of cellular processes, such as cell death, and infer various cellular states. The provided results show the inherent separability of distinct cell death processes. In future work, the proposed model could be further refined to solve for the optimal point of detection of differentiation and similar models could be extended to other cellular processes or genomic states.

CHAPTER 7

FUTURE WORK

7.1 Hyperspectral Histopathological Image Modalities

This thesis has investigated several imaging modalities for cancer, but many others exist. A particular category of modality that would be important to explore in future research is hyperspectral imaging. In applications ranging from environmental monitoring to biomedical imaging, it is useful to acquire absorbance and reflectance measurements of light beyond the visible spectrum. In biomedical applications, absorption of different spectra of light is indicative of the chemical composition of the imaged sample and can provide quantitative measurements of cells and tissue in a *label-free* manner, obviating the need for stains that contaminate the specimen. These quantitative measurements can supplement H&E stained samples in distinguishing different cell types, which is ultimately of use to pathologists in detecting cancer. Fourier transform infrared (FTIR) spectroscopy and hyperspectral fluorescence lifetime imaging (FLIM) are two useful spectral imaging techniques for cancer. An example intensity band of FTIR from a breast cancer dataset is shown in Fig. 7.1, and the distribution of frequencies of FTIR absorbance spectra for a dataset of breast cancer patients for different cell types is shown in Fig. 7.2. Example FLIM spectra for several different tissue types are shown in Fig. 7.3.

Extracting spectra from hyperspectral data and analyzing the recorded spectra poses unique challenges. Due to insufficient imaging resolution and the heterogeneity of samples, pixels in a hyperspectral data volume are likely to contain spectral information of several cells. Classical clustering algorithms, such as Gaussian mixture models (GMM) or K-means, attempt to classify the spectrum at each pixel as belonging predominantly to a particular class of cell and do not account for the mixing of cell types in the acquisition

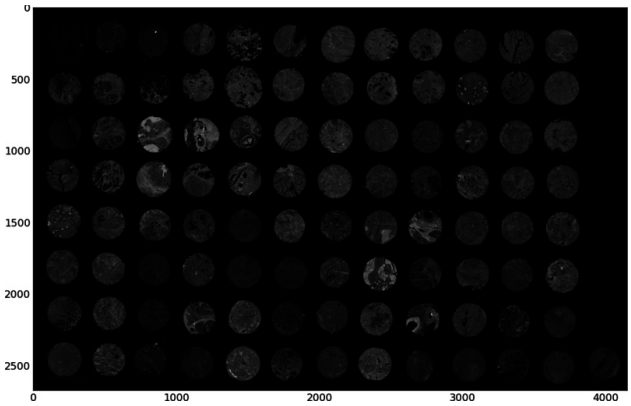
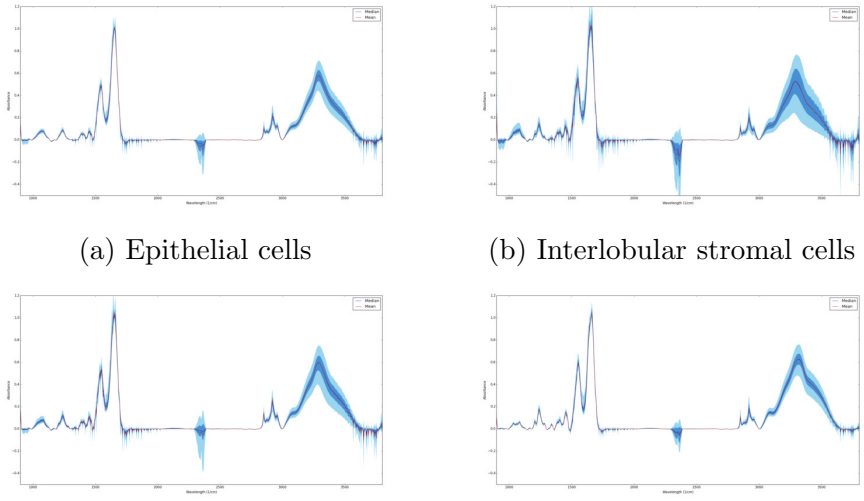


Figure 7.1: Intensity of a sample frequency band from a dataset of breast cancer biopsies imaged with FTIR.



(a) Epithelial cells

(b) Interlobular stromal cells

(c) Myoepithelial cells

(d) Stromal cells

Figure 7.2: Mean and spread of FTIR absorbance spectra of cells in breast cancer tumors. Absorbance was sampled at frequencies from 850 μm to 1850 μm at an interval of 2 μm.

process.

Clustering analysis run on the FLIM spectra, with its corresponding H&E slice, is shown in Fig. 7.4. Several iterations of both GMM and K-means were run, to account for variation in the random initialization, and the qualitatively best resulting clustering assignment is shown. For both algorithms, seven components, or clusters, were assumed. Experiments were conducted with a varying number of clusters, but seven seemed to produce the best compromise of generalization and specificity. Both the GMM and K-means

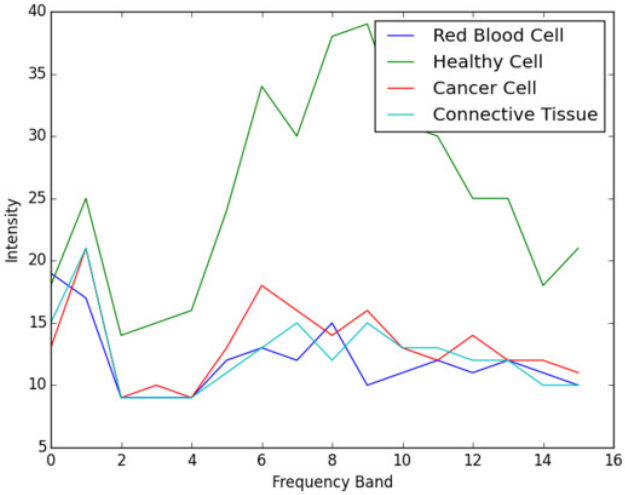
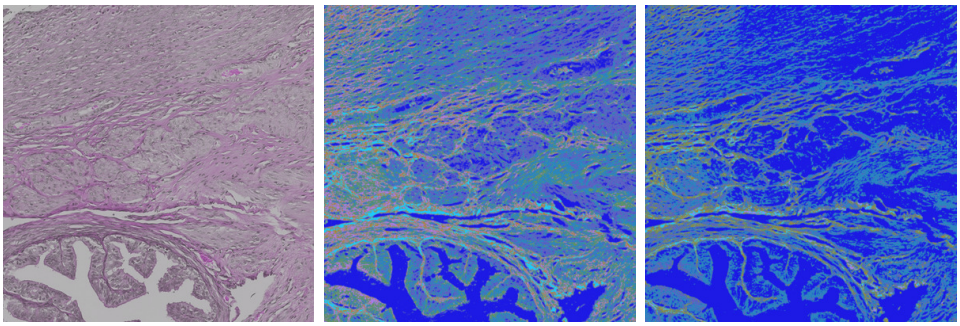


Figure 7.3: Example hyperspectral FLIM spectra of cancerous ovary tissue.

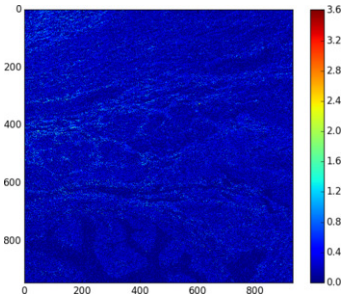


(a) H&E (b) GMM (c) K-means

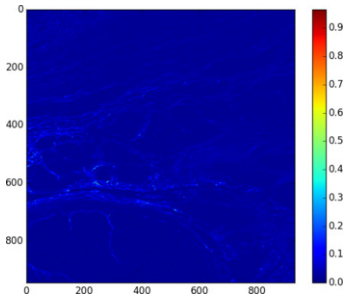
Figure 7.4: Class labels of ovary data generated by GMM and K-means clustering methods using 7 components.

results show a change in cluster assignment moving from left to right across the image, which is a result of artifacts in the mosaicing process. However, K-means seems to perform poorly, as it incorrectly classifies many pixels in the image as belonging to the background.

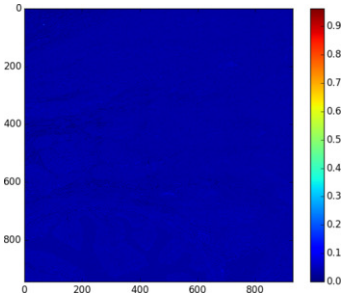
The same linear unmixing tools mentioned previously for gene expression would be appropriate for hyperspectral imaging, and in fact were developed for such imaging. Linear unmixing was performed on the same FLIM data using the popular N-FINDR algorithm [58], and the resulting abundance maps of each of the seven discovered endmembers are shown in Fig. 7.5. The presence of several endmembers correlates significantly with the locations of specific cell types in the image. One cell type of note is red blood cells, which



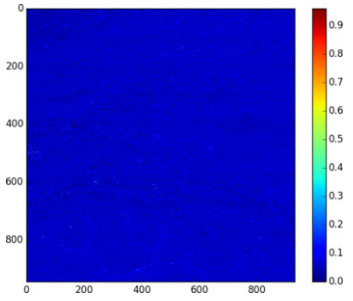
(a) Abundance of endmember 1.



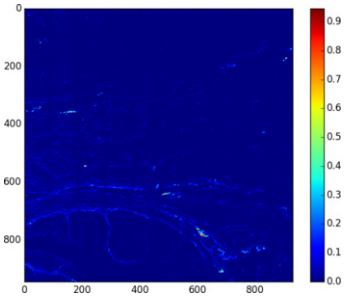
(b) Abundance of endmember 2.



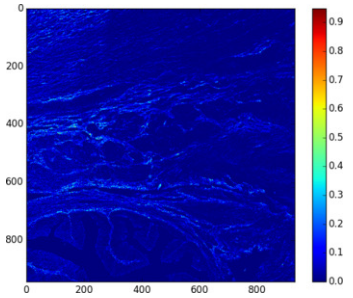
(c) Abundance of endmember 3.



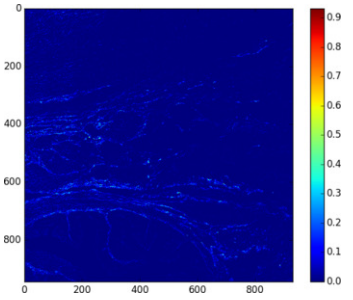
(d) Abundance of endmember 4.



(e) Abundance of endmember 5.

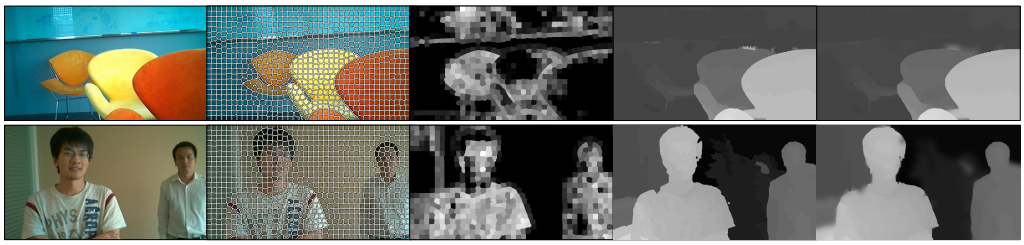


(f) Abundance of endmember 6.



(g) Abundance of endmember 7.

Figure 7.5: Abundance maps of discovered endmembers from an ovary tissue slide using linear unmixing. The intensity at a pixel in each map corresponds to the contribution of that endmember to the spectrum at that pixel in the hyperspectral FLIM image. Endmembers 1, 3, and 4 are present nearly uniformly throughout the sample. Endmember 5 corresponds to red blood cells. Endmembers 6 and 7 seem to be present mostly in connective tissue.



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

Figure 7.6: Result on two test images (stereo pairs not shown) from each step of depth image acquisition. (a) Original images. (b) Segmentation images from the SLIC algorithm. (c) Edge density images. (d) Depth images after aggregation. (e) Depth images after refinement.

appear strongly in the abundance map of the fifth endmember. Interestingly, blood cells were also the easiest cells to detect using K-means clustering in the FTIR data.

Future work could leverage the same H&E pipeline discussed in this thesis for extracting image features of H&E data, which could then be connected to different representations of hyperspectral data of adjacent slices. Additionally, future work could integrate hyperspectral data into an imaging-genomics framework for providing insights into the connections between genomic data and phenotype that cannot be captured through H&E staining.

7.2 Graphical Image Representation for Inference

In addition to the grading of pleomorphism of nuclei, based on characteristics of shape, texture, and intensity, a pathologist will also inspect the spatial structure of a cancerous tumor when giving a diagnosis and prognosis. An example metric that may be considered is the quantification of tumor-infiltrating lymphocytes, which has been shown to be indicative of survival in breast cancer patients [6]. Graphical models are the most comprehensive tools available for describing spatial relationships, but often, simple metrics quantifying the heterogeneity of local patches have been used to date in computational pathology research.

Sophisticated graphical models have been developed in other image processing problems that could be applied in future research to cancer imaging. An example of such an image processing problem is that of stereo matching,

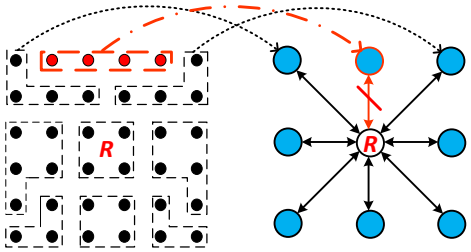
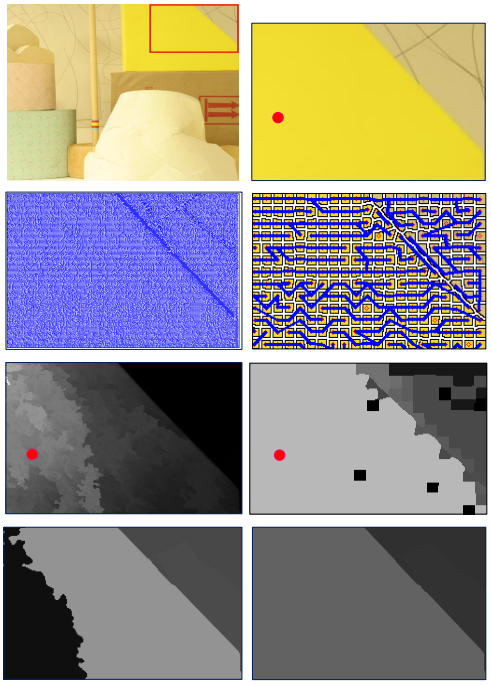


Figure 7.7: Region graph builder. Superpixels resulting from the SLIC over-segmentation method are treated as nodes on an eight-connected undirected regular graph. The red colored superpixel does not share any neighbor pixel with superpixel R. Therefore the edge between them is penalized.



(a) (b)

Figure 7.8: Examination of results from proposed hierarchical minimum spanning trees approach. Row 1: (a) Original lamp shade image. (b) A close-up patch from the original image with the red test point. Row 2: (a) Pixel-level MST. (b) Region-level MST. Row 3: (a) The weighted contribution of all pixels inside the close-up patch to the test point according to the pixel-level MST. (b) The contribution according to the region-level MST. Row 4: (a) Depth image result using the pixel-level MST only. (b) Result using the proposed adaptive fusion of pixel-level and region-level MSTs.

in which the depth of pixels in an image are inferred from the image and its stereo pair. This problem is most comprehensively solved using graphical models, which can enforce a smoothness constraint upon the inferred depth image, capturing the intuition that neighboring pixels should be at a similar depth from the camera.

An algorithm for solving this problem in a computationally efficient, yet highly accurate, manner using hierarchical minimum spanning trees (MST) was developed in a previous work [66]. Similar to H&E analysis, the image is first segmented, in this case using SLIC superpixels, as shown in Fig. 7.6. Next, two graphs $\left\{G_P = (V^{(P)}, E^{(P)}), G_S = (V^{(S)}, E^{(S)})\right\}$ are defined on the image: one at the pixel-level P, where each pixel is a node V with edges E between nodes; and one the region-level, or superpixel-level, S, respectively. Weights ω are defined between nodes in each graph based on their color similarity:

$$\omega_P(p, q) = |I(p) - I(q)|, \tag{7.1}$$

$$\omega_R(S, T) = |I_S - I_T|. \tag{7.2}$$

From these weights, a minimum spanning tree T is defined for each graph. Defining neighboring nodes on the superpixel MST is accomplished as shown in Fig. 7.7. Then MST's are built on both the pixel and superpixel levels, as shown in Fig. 7.8.

The goal of the algorithm is to find the disparity, which is related to the pixel's depth, at each pixel between the two stereo images. The disparity is the distance between a pixel in one image and its corresponding point in the stereo pair. To determine the correspondence between pixels across images, the algorithm searches for the pixels or nodes that appear similar in the images. A cost function is defined to measure this similarity. For this algorithm, the cost $C_d(V_i)$ for a given distance d at a node V_i is defined by the color similarity of the node in one image and its corresponding node in the paired image. To enforce spatial consistency, this cost is aggregated over nodes of the graph, weighted by the edge weights ω along the path $P(V_i, V_j)$ to each other node V_j,

$$C_d^A(V_i) = \sum_{V_j \in T} \exp\left(-\frac{\sum_{\omega_E \in P(V_i, V_j)} \omega_E}{\sigma}\right) C_d(V_j).$$

The exponential function with parameter σ tapers the influence of nodes that are farther from the considered node V_i. Finally, the cost at both the pixel and region levels are aggregated and mixed together, forming a final cost,

$$C_d'^A(p) = \alpha_R C_d^A(p) + (1 - \alpha_R) C_d^A(R),$$

where α_R is a mixing parameter.

The benefit of using this hierarchical structure of MSTs can be seen from the inferred depth images in Fig. 7.6, especially when compared to the result using a pixel-level tree only, as shown in Fig. 7.8. Similar graph based techniques for inference could be used in WSIs to detect regions of tumors, such as epithelial and stromal, or structures such as tubules, in an efficient manner, which will be crucial in the future for more accurate and computationally efficient prognosis.

REFERENCES

[1] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–74, 2011.

[2] TCGA Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012.

[3] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Graf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerod, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Borresen-Dale, J. D. Brenton, S. Tavare, C. Caldas, and S. Aparicio, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.

[4] C. C. Jaffe, "Imaging and genomics: Is there a synergy?" *Radiology*, vol. 264, no. 2, pp. 329–331, 2012.

[5] Y. Yuan, H. Failmezger, O. M. Rueda, H. R. Ali, S. Gräf, S.-f. 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*, vol. 4, no. 157, pp. 157ra143—-157ra143, 2012.

[6] H. R. Ali, E. Provenzano, S. Dawson, F. M. Blows, B. Liu, M. Shah, H. M. Earl, C. J. Poole, L. Hiller, J. A. Dunn, S. J. Bowden, C. Twelves, J. M. S. Bartlett, S. M. A. Mahmoud, E. Rakha, I. O. Ellis, S. Liu, D. Gao, T. O. Nielsen, P. D. P. Pharoah, and C. Caldas, "Association between CD8 + T-cell in filtration and breast cancer survival in 12 439 patients," *Annals of Oncology*, vol. 25, no. June, pp. 1536–1543, 2014.

[7] C. C. Maley, K. Koelble, R. Natrajan, A. Aktipis, and Y. Yuan, "An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer," *Breast Cancer Research*, vol. 17, no. 131, 2015.

[8] S. Nawaz, A. Heindl, K. Koelble, and Y. Yuan, "Beyond immune density: Critical role of spatial heterogeneity in estrogen receptor-negative breast cancer," *Modern Pathology*, vol. 28, no. 6, pp. 766–777, 2015.

[9] R. Natrajan, H. Sailem, F. K. Mardakheh, M. A. Garcia, C. J. Tape, M. Dowsett, C. Bakal, and Y. Yuan, "Microenvironmental heterogeneity parallels breast cancer progression: A histologygenomic integration analysis," *PLoS Medicine*, vol. 13, no. 2, 2016.

[10] C. Wang, T. Pécot, D. L. Zynger, R. Machiraju, C. L. Shapiro, and K. Huang, "Identifying survival associated morphological features of triple negative breast cancer using multiple datasets." *Journal of the American Medical Informatics Association*, vol. 20, no. 4, pp. 680–7, 2013.

[11] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, 2010.

[12] L. A. D. Cooper, J. Kong, F. Wang, T. Kurc, C. S. Moreno, D. J. Brat, and J. H. Saltz, "Morphological signatures and genomic correlates in glioblastoma," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, pp. 1624–1627.

[13] J. Kong, L. A. D. Cooper, F. Wang, D. A. Gutman, J. Gao, C. Chisolm, A. Sharma, T. Pan, E. G. Van Meir, T. M. Kurc, C. S. Moreno, J. H. Saltz, and D. J. Brat, "Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 12 PART 2, pp. 3469–3474, 2011.

[14] 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," *Laboratory Investigation*, vol. 95, no. 4, pp. 366–376, 2015.

[15] L. A. D. Cooper, A. B. Carter, A. B. Farris, F. Wang, J. Kong, D. A. Gutman, P. Widener, T. C. Pan, S. R. Cholleti, A. Sharma, T. M. Kurc, D. J. Brat, and J. H. Saltz, "Digital pathology: Data-intensive frontier in medical imaging," *Proceedings of the IEEE*, vol. 100, no. 4, pp. 991–1003, 2012.

[16] R. Colen, I. Foster, R. Gatenby, M. E. Giger, R. Gillies, D. Gutman, M. Heller, R. Jain, A. Madabhushi, S. Madhavan, S. Napel, A. Rao, J. Saltz, J. Tatum, R. Verhaak, and G. Whitman, "NCI Workshop Report: Clinical and computational requirements for correlating imaging phenotypes with genomics signatures," *Translational Oncology*, vol. 7, no. 5, pp. 556–569, 2014.

[17] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "CellProfiler: Image analysis software for identifying and quantifying cell phenotypes," *Genome Biology*, vol. 7, no. 10, p. R100, 2006.

[18] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. a. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. a. Akslen, O. Fluge, a. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, a. L. Børresen-Dale, P. O. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[19] 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*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011.

[20] D. C. Fernandez, R. Bhargava, S. M. Hewitt, and I. W. Levin, "Infrared spectroscopic imaging for histopathologic recognition," *Nature Biotechnology*, vol. 23, no. 4, pp. 469–474, 2005.

[21] F. N. Pounder, K. Reddy, and R. Bhargava, "Development of a practical spatial-spectral analysis protocol for breast histopathology using Fourier transform infrared spectroscopic imaging," *Faraday Discussions*, vol. 187, pp. 43–68, 2016.

[22] P. D. E. Beule, D. M. Owen, H. B. Manning, C. B. Talbot, J. Requejo-isidro, C. Dunsby, J. M. C. Ginty, R. K. P. Benninger, D. S. Elson, I. A. N. Munro, M. J. Lever, P. Anand, M. A. A. Neil, and P. M. W. French, "Rapid hyperspectral fluorescence lifetime imaging," *Microscopy Research and Technique*, vol. 70, no. 5, pp. 481–484, 2007.

[23] A. J. Bower, B. Chidester, J. Li, Y. Zhao, M. Marjanovic, E. J. Chaney, M. N. Do, and S. A. Boppart, "A quantitative framework for the analysis of multimodal optical microscopy images," *Quantitative Imaging in Medicine and Surgery*, vol. 7, no. 1, pp. 24–37, 2017.

[24] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009.

[25] M. Veta, J. P. W. Pluim, P. J. V. Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1400–1411, 2014.

[26] J. A. A. Jothi and V. M. A. Rajam, "A survey on automated cancer diagnosis from histopathology images," *Artificial Intelligence Review*, pp. 1–51, 2016.

[27] 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*, vol. 7, pp. 97–114, 2014.

[28] D. L. Rubin, K.-h. Yu, C. Zhang, G. J. Berry, R. B. Altman, C. Re, and M. Snyder, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature communications*, vol. 7, no. 12474, 2016.

[29] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quanititative analysis," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 1107–1110.

[30] J. H. Phan, C. F. Quo, C. Cheng, and M. D. Wang, "Multiscale integration of -omic, imaging, and clinical data in biomedical informatics," *IEEE Reviews in Biomedical Engineering*, vol. 5, pp. 74–87, 2012.

[31] S. Kothari, J. H. Phan, and M. D. Wang, "Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade," *Journal of Pathology Informatics*, vol. 4, p. 22, 2013.

[32] M. Veta, A. Huisman, M. A. Viergever, P. J. V. Diest, and J. P. W. Pluim, "Marker-controlled watershed segmentation of nuclei in H&E stained breast cancer biopsy images," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, pp. 618–621.

[33] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[34] 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*, vol. 12, no. 484, 2011.

[35] L. A. D. Cooper, J. Kong, D. A. Gutman, F. Wang, J. Gao, C. Appin, S. Cholleti, T. Pan, A. Sharma, L. Scarpace, T. Mikkelsen, T. Kurc, C. S. Moreno, D. J. Brat, and J. H. Saltz, "Integrated morphologic analysis for the identification and characterization of disease subtypes," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 317–323, 2012.

[36] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art uperpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2281, 2012.

[37] L. E. Boucheron, "Object- and spatial-Level quantitative analysis of multispectral histopathology images for detection and characterization of cancer," Ph.D. dissertation, University of California Santa Barbara, 2008.

[38] L. A. D. Cooper, J. Kong, D. A. Gutman, F. Wang, S. R. Cholleti, T. C. Pan, P. M. Widener, A. Sharma, T. Mikkelsen, A. E. Flanders, D. L. Rubin, E. G. Van Meir, T. M. Kurc, C. S. Moreno, D. J. Brat, and J. H. Saltz, "An integrative approach for in silico glioma research," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2617–2621, 2010.

[39] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer-Verlag, 2009.

[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. GA: USENIX Association, 2016. pp. 265–283.

[41] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[42] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[43] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 1106–1114.

[44] K. Sirinukunwattana, S. E. A. Raza, Y.-w. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[45] J. Xu, L. Xiang, Q. Liu, S. Member, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 119–130, 2016.

[46] F. Xing, Y. Xie, and L. Yang, "An automatic learning-based framework for robust nucleus segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 550–566, 2016.

[47] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, no. 29, 2016.

[48] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, "AggNet : Deep learning from crowds for mitosis detection in breast cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1313–1321, 2016.

[49] E. D. Gelasca, B. Obara, D. Fedorov, K. Kvilekval, and B. S. Manjunath, "A biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC Bioinformatics*, vol. 10, no. 368, 2009.

[50] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[51] J. Quackenbush, "Microarray analysis and tumor classification," *New England Journal of Medicine*, vol. 354, no. 23, pp. 2463–2472, 2006.

[52] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[53] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[54] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[55] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. Mcdermott, and P. J. Campbell, "COSMIC: Exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Research*, vol. 43, no. October 2014, pp. 805–811, 2015.

[56] C. Bazot, N. Dobigeon, J.-y. Tourneret, A. K. Zaas, G. S. Ginsburg, and A. O. H. III, "Unsupervised Bayesian linear unmixing of gene expression microarrays," *BMC Bioinformatics*, vol. 14, no. 99, 2013.

[57] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[58] M. E. Winter, "N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data,," in *SPIE Conference on Imaging Spectrometry V*, vol. 3753, no. July, 1999, pp. 266–275.

[59] O. Eches, N. Dobigeon, C. Mailhes, and J.-Y. Tourneret, "Bayesian estimation of linear mixtures using the normal compositional model. Application to hyperspectral imagery," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1403–1413, 2010.

[60] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2008.

[61] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, 2011.

[62] Y. Zhao, M. Marjanovic, E. J. Chaney, B. W. Graf, Z. Mahmassani, M. D. Boppart, and S. A. Boppart, "Longitudinal label-free tracking of cell death dynamics in living engineered human skin tissue with a multimodal microscope," *Biomedical Optics Express*, vol. 5, no. 10, p. 3699, 2014.

[63] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "Nonrigid multimodality image registration," *Proceedings of SPIE*, vol. 4322, pp. 1609–1620, 2001.

[64] M. Styner, S. Member, C. Brechbühler, G. Székely, and G. Gerig, "Parametric estimate of intensity inhomogeneities applied to MRI," *IEEE Transactions on Medical Imaging*, vol. 19, no. 3, pp. 153–165, 2000.

[65] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[66] D. T. Vu, B. Chidester, H. Yang, M. N. Do, and J. Lu, "Efficient hybrid tree-based stereo matching with applications to postcapture image refocusing," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3428–3442, 2014.