Thesis

Automated Detection of Circulating Cells Using Low Level Features

Submitted by

Tegan Halley Emerson

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2013

Master's Committee:

Advisor: Michael Kirby

Chris Peterson
Jennifer Nyborg

Abstract

Automated Detection of Circulating Cells Using Low Level Features

This thesis addresses the problem of detection of high definition circulating tumor cells using data driven feature selection. We propose techniques in pattern analysis and computer vision to achieve this goal. Specifically, we determine a set of low level features which can structurally differentiate between different cell types of interest to contribute to the treatment and monitoring of patients. We have implemented three image representation techniques on a curated data set. The curated data set consists of digitized images of 1000 single cells: 500 of which are high definition circulating tumor cells or other cells of high interest, and 500 of which are white blood cells. None of the three image representation techniques have been previously applied to this data set. One image representation is a novel contribution and is based on the characterization of a cell in terms of its concentric Fourier rings. The Fourier Ring Descriptors (FRDs) exploit the size variations and morphological differences between events of high and low interest while being rotationally invariant. Using the low level descriptors, FRDs, as a representation with a linear support vector machine decision tree classifier we have been able to average 99.34% accuracy on the curated data set and 99.53% on non-curated data. FRDs exhibit robustness to rotation and segmentation error. We discuss the applications of the results to clinical use in context of data provided by The Kuhn Laboratory at The Scripps Research Institute.

Acknowledgements

I would like to acknowledge the incredible support and encouragement of my advisor Dr. Michael Kirby in pursuing this project for my master's work. I would also like to thank my research collaborators Dr. Peter Kuhn, Dr. Anand Kolatkar, and Dr. Mohsen Sabouri at The Kuhn Laboratory for taking a chance on me, for so patiently working with me as I began to learn the biological side of our project, and for being sounding boards and contributing to the development of the methods discussed. Additionally, I would like to thank Dr. Paul Newton from the University of Southern California Viterbi School of Engineering for contributing to the development of many of the ideas that follow as well as helping to teach me how to bridge the language barrier between mathematics and biology. Also, I would like to thank my committee members Dr. Chris Peterson and Dr. Jennifer Nyborg for being a part of this work. I would like to acknowledge my family members, past and present, who provided the desire to work on a project concerning cancer: Marlene Ostrer, Robert Ostrer, Margret Emerson, and Kenneth Emerson. Lastly, I would like to thank my parents, Dawn and Bruce Emerson, my brother, Skye Emerson, and my close friends, Meghan Kahnle and Kate Willis, for providing, from a distance, the emotional support to make this possible. I would like to dedicate this work to Marlene Ostrer who has survived breast cancer twice and Kenneth Emerson who passed away after a several year battle with metastatic prostate cancer.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Problem Statement and Other Related Research

This thesis concerns the application of pattern analysis and computer vision techniques to the task of computer aided diagnostics in the realm of circulating tumor cells. We propose to determine structurally differentiating features which aid in detection of cell populations of interest. All features are extracted from images generated using the high definition circulating tumor cell assay developed by The Kuhn Laboratory (TKL) at The Scripps Research Institute (TSRI). High definition circulating tumor cells (HD-CTCs) are cells in the blood that originated from tumors of epithelial origin. Research surrounding circulating tumor cells has been inspired by a desire for greater understanding of the metastatic phase of cancer as well as a hope for earlier stage diagnosis and better monitoring of patient status and treatment [1]. The term HD-CTC is unique to the assay which has been developed by TKL at TSRI.

Circulating tumor cells (CTCs) are disseminated tumor cells in the blood. In 2011 it was asserted that the "CTCs represent an independent predictor of outcome in patients with metastatic breast cancer." [2] In addition, similar connections have been shown, in the literature, between CTCs and other cancers including metastatic colorectal cancer, neuroendocrine tumors, non-small cell lung cancer, and progressive castration-resistant prostate cancer [3, 4, 5, 6]. There is evidence to suggest that by better understanding CTCs we will have a better understanding of cancer in its metastatic phase. It has been demonstarted that a large increase in CTCs in a patient after receiving chemotherapy is a strong predictor for relapse and acts, in some ways, as a marker for the aggressiveness of the tumor [7]. Thus, there is reason to believe that greater understanding of CTCs will aid in the monitoring of

patients as their cancer progresses. Additionally, there is hope that CTCs will eventually be able to be used to help in the detection of earlier stage cancer which results in higher survival rates.

While strong connections have been made between the relative counts of CTCs in a patient blood sample and prognosis in many types of cancer, there is only one Food and Drug Administration (FDA) approved technique for the automated detection of circulating tumor cells called Cell Search [8]. Cell Search uses positive enrichment based on epithelial cell adhesion molecule (EpCAM) and then uses cytokeratin to mark for CTCs within the reduced population. This method, however, has been shown to produce lower CTC counts than the assay used at TKL [1]. It is believed that the difference in CTC counts is largely caused by undergoing enrichment prior to cytokeratin staining. The downside, however, to the current method at TKL is that it requires manual involvement of a highly trained technician to produce these counts and that it inherits the non-standardized results associated with human involvement.

We thus seek to detect CTCs found in the TKL assay by determining a set of low level features which strongly differentiates between cell populations. This problem has two primary parts. First, to improve the efficiency and standardization of detection of rare events of high interest through low level features. Second, to determine unsupervised subclasses of cell types of both high and low interest. The work presented in [9] addressed the results of first efforts on the second problem. This thesis, however, focuses on the detection and classification of known classes. It is our hope that the results of this work, and future work, will contribute to the treatment and monitoring of patients with diseases with epithelial cell expression.

We are the first group to focus on the data driven feature selection detection of circulating tumor cells using the TKL high definition circulating tumor cell assay. There have been, however, others who have applied computer vision techniques to the subclassification of human epithelial cells. The research done on classification of epithelial cell subtypes will be of high relevance to us in the subclassification of events of high interest because high definition circulating tumor cells are tumor cells that came from a tumor of epithelial origin, and are, consequently, epithelial cells themselves. Due to the number of diseases that have epithelial cell expression, research on human epithelial cells has grown significantly. Much of the research, however, strongly differs from our own.

In 2012 there was a contest on classification of types of human cell line epithelial cells (HEp-2) hosted by the International Conference on Pattern Recognition. First, the classification task outlined in the competition was to classify an image containing multiple cells (all of which are epithelial cells) as one of several subtypes based on the dominant cell subtype present. In the contest data challenge, researchers are not first asked to distinguish the epithelial cells from other cells in the images. Additionally, all images contained in the data set are of high magnification, focus on the nucleus of the cell, and contestants are provided with masks for each cell contained in the image. There were 28 teams that participated in the competition. A description of each teams methodology can be found by looking at the results page of the contest website [10].

We, unlike the contestants, must build a classifier that can first distinguish events of interest from white blood cells, then additional classifiers to subdivide our populations of interest. Furthermore, in observing the contest data one can see an obvious distinction between the cell classes and that the imaging shows a large level of textural variation in

the nucleus of the cells. This textural variation allows ready application of many patch based feature extraction, many of which are discussed in [10]. Thus, while there have been selected attempts at classifying human epithelial cells, none to our knowledge seek to both identify these events in human blood in the presence of white blood cells and determine subclassifications of them from low magnification images and multiple image channels.

We will first discuss the data we have analyzed, the current methods used for detection, and some of the challenges associated with the data. Next, we will describe the three different feature extraction and image representations we have applied to the data followed by a discussion of the classification technique we have used. Lastly, we will present the results of our work to current date and discuss future ambitions related to this project.

CHAPTER 2

The Data

The data used in this thesis was generated by The Kuhn Laboratory (TKL) at The Scripps Research Institute (TSRI). [1] In this chapter we describe the data acquisition process, challenges of the data, and the manual curation of a small data set.

2.1. Acquisition of Data

2.1.1. Imaging. The process begins with a blood draw performed on a patient of interest. Next, the blood is placed in a centrifuge and all red blood cells are removed. The remaining blood is then slided and treated with stains to mark for differentiating characteristics of cells. There are three fluorescent stains common to all patient samples in the HD-CTC assay: a nuclear marker 4',6-diamidino-2-phenylindole ($DAPI$), a cytokeratin (CK) marker called $Alexa555$, and $Alexa647$ which is a marker for protein tyrosine phosphatase receptor type C (CD-45). In some cases more stains are used depending on the nature of the disease of the patient. For example, in the case of prostate cancer a fourth stain is used to mark for an androgen receptor in addition to the three common stains. After storing the slides for a specified amount of time, the slides are thawed and the treated blood is imaged using immunofluorescent microscopy at the excitation wavelength of each stain. A 10X objective on an inverted microscope is used to generate the images. The results of the imaging process are three monochromatic images corresponding to the intensities from each of the emission wavelengths. The excitation and emission wavelengths are 555nm and 565nm respectively for $Alexa555$, 650nm and 668nm respectively for $Alexa647$, and 350nm and 470nm respectively

[1]This center is a part of collaborative network of 12 Physical Science-Oncology Centers (PS-OCs), funded by The National Cancer Institute (NCI) NCI Grant Number: U54CA143906

for $DAPI$. The three monochromatic images are then layered into a false colored image where the red layer corresponds to the $Alexa555$ channel, the green layer corresponds to the $Alexa647$ channel, and the blue layer corresponds to the $DAPI$ channel. A description of the imaging process, in greater detail, can be found in [1].

For each slide there are 2,304 images generated for each channel combining to produce 6,912 total images per slide. Each slide contains approximately three million events, of which less than 0.1% are truly HD-CTCs. One of the distinguishing traits of the HD-CTC assay is that there is no enrichment prior to imaging. For this reason there are more than one type of cell of high interest to detect in a patient sample.

2.1.2. The TKL Algorithm. The algorithm for detection of these rare cell events currently in use by TKL can be broken into two stages. The first stage is computer automated while the second stage is manual and done by a hematopathologist-trained technical analyst.

Segmentation to determine the location of all cell events is performed on the monochrome $DAPI$ channel image using software called ImageJ. One benefit to performing segmentation on the $DAPI$ channel is that it is typically the cleanest of the stains leaving very few artifacts. Once an event has been detected a center of mass computation is performed to identify the center of the cell. After the center of the cell has been determined two circles of fixed radii, with the cell center as the circle center, are computed. The larger circle of radius 40 pixels is overlaid in the $Alexa555$ channel and the average image value within the circle, to be referred to as α, is computed. Next, the smaller circle with a radius of 14 pixels is overlaid in the $Alexa647$ channel and the average image value, to be referred to as β, within this circle is computed.

A table is then generated for a slide with CK intensity along the columns decreasing from left to right and with CD-45 intensity along the rows increasing from top to bottom. The intensities are computed using a quantity determined by the number of standard deviations over the mean (SDOM). For each image an average image value, $\bar{A}_{channel}$, and standard deviation, $\sigma_{channel}$ is computed for both $Alexa555$ or $Alexa647$ channels. The SDOM measurement of α is given by

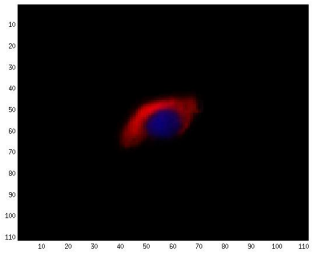
$$\tilde{\alpha} = \frac{\alpha - \bar{A}_{Alexa555}}{\sigma_{Alexa555}}.$$

Similarly, an SDOM measurement is computed for β as

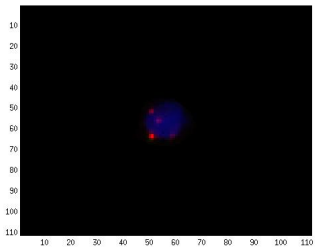
$$\tilde{\beta} = \frac{\beta - \bar{A}_{Alexa647}}{\sigma_{Alexa647}}.$$

The table generated identifies events deemed to be candidate HD-CTCs as well as their location and cropped images of the event in each invididual channel and a cropped composite image. The format of the table places the events such that the events the algorithm believes to be the most likely HD-CTCs are listed first and as you move through the table the likelihood of the events being HD-CTCs reduces.

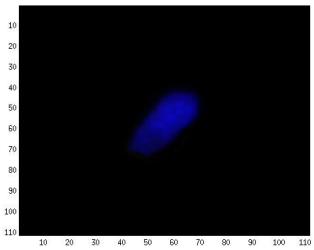
It is at this point that the automatically generated table is passed to a technician to manually classify each candidate event as either an HD-CTC, one of three near HD-CTC types (Apoptotic, No CK, Small CK), imaging noise, or as an event of non-interest (e.g., a WBC). Examples of each type of manual classification type can be seen in Figure 2.1. Classification of an event as a HD-CTC requires that $\tilde{\alpha}$ is generally greater than or equal to 6 while also presenting in the lowest 0.2 % of the population in the $Alexa647$ channel together with a sufficiently large nucleus, in addition to other visually discernable characteristics. A classification of an event as "No CK" arises when the size of the nucleus of an event meets



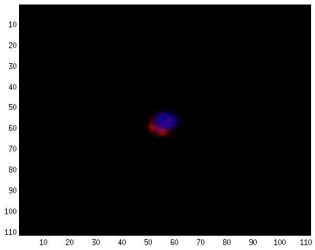
(a) HD-CTC



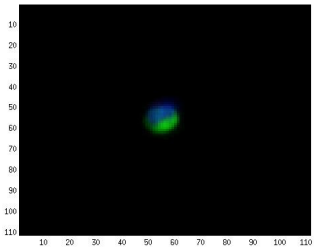
(b) Apoptotic



(c) No CK



(d) Small CK



(e) WBC

FIGURE 2.1. The set of images above are sample cells from the curated data set of each class of cell we seek to identify.

HD-CTC	Apoptotic	No CK	Small CK
Large Nucleus	Sick/ Unhealthy Cell	Nucleus Sufficiently Large to be HD-CTC	Nucleus Too Small to be an HD-CTC
$\tilde{\alpha} > 6$ High Cytokeratin Expression (typically)	No Definitive Size Threshold	Low or Non-existent Cytokeratin	$\tilde{\alpha} > 6$ High Cytokeratin Expression
Cytokeratin Pocket for Nucleus	Often Small with Cytokeratin Speckles	No Cytokeratin Pocket for Nucleus	Cytokeratin Pocket for Nucleus
CD-45 Expression in Bottom 2%	CD-45 Expression in Bottom 2%	CD-45 Expression in Bottom 2%	CD-45 Expression in Bottom 2%

FIGURE 2.2. This figure contains a comparison of the properties of each cell type of high interest.

the accepted criterion of an HD-CTC but does not exhibit appropriate levels of CK. "Small CK" refers to an event that has an $\tilde{\alpha}$ that is sufficiently high to be considered an HD-CTC, as well as the some of the visually discernable characteristics, but an insufficient nuclear size. Lastly, "Apoptotic" refers to a set of cells that do not fall directly into the category of HD-CTC, No CK, or Small CK, but share some chanracteristics of each population and appear to be sick or dying. A comparison of the properties of each cell type are summarized in Figure 2.2. Currently, once a technician has reached a point in the table where they have classified 200 events in a row as WBCs, they terminate manual classification. We aim to produce a smaller set of candidate events for a technician to manually classify using our data driven features while still including all HD-CTC and near HD-CTC types and minimizing events of non-interest.

2.2. Challenges of the Data

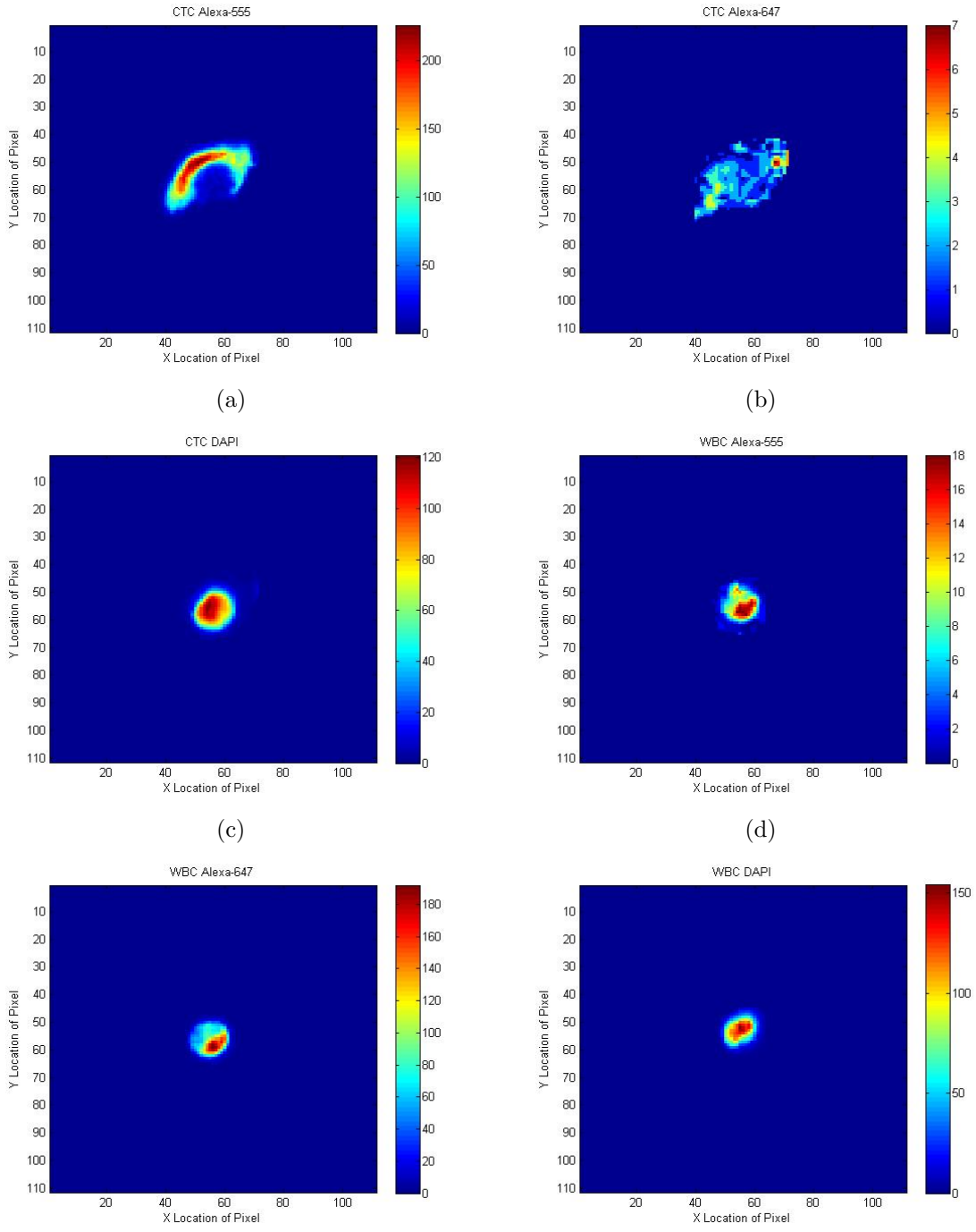
2.2.1. Rare Event Detection. Rare event detection problems arise when the target events occur at significantly lower frequencies than non-target events. "Lower frequencies" often refers to orders of magnitude differences in the literature, often on the scale of 1 target event per 1 thousand non-target events. Examples of non-medical rare event detection problems include detection of chemical particles in air from hyperspectral images and detection of rare plant species from hyperspectral images. In recent years there has been an increase in application of rare event detection to various biomedical queries. Many of these applications focus on detection of cells of interest in various mediums including blood, tissue, and bone marrow [11, 12, 13, 14, 15]. The research to be discussed falls into this category as we are seeking to identify HD-CTCs as well as other cells of high interest which occur on a scale of 5-1000 per 3 million cells in an image set generated from one slide of patient blood.

2.2.2. Image Quality and Characteristics. Our algorithms are currently being run on 8-bit Joint Photographic Experts Group (JPEG) images generated using a 10x objective which is automatically focused. While the images are generated using high-end imaging techniques there is still variation in the image quality due to the automated focusing of the objective as well as artifact signatures based on who was responsible for the preparation of the slide. Furthermore, at 10x scale there is very limited textural information contained in the images which is an issue for image classification tasks [16].

Figure 2.3 illustrates the lack of textural variation in an event of interest and of low interest. Additionally, due to the biological properties of the different cells we see significant variation in the expression of the three different markers on different cells within an image. We also see inter-slide variation in expression that can be attributed to small concentration variations in the stains. It has been suggested by members of TKL that dimness may in fact be related to an unsupervised cell subtype [17], and until further analysis has been performed we assume that the variation in cell expression may contain relevant information.

2.2.3. High Throughput. Based on the current assay there are 2,304 images generated per slide in each channel resulting in 6,912 total images per slide. However, a full exploration of a patient's CTC volume is judged based on a four slide sample when possible. Thus, for a comprehensive patient evaluation there are 27,648 total images to draw information from and roughly 12 million events. In order to maintain a manageable workflow we need to be able to produce all candidate cells for a single slide in approximately one hour with satisfactory sensitivity and specificity.[2] Sensitivity refers to our ability to detect all the events of high interest and specificity refers to the relative amount of extraneous events we pass to the

[2]The current algorithm used by the laboratory can produce a set of candidate cells in approximately one hour. Our sensitivity and specificity will be measured in comparison to the the set of candidate cells currently produced using the current algorithm.

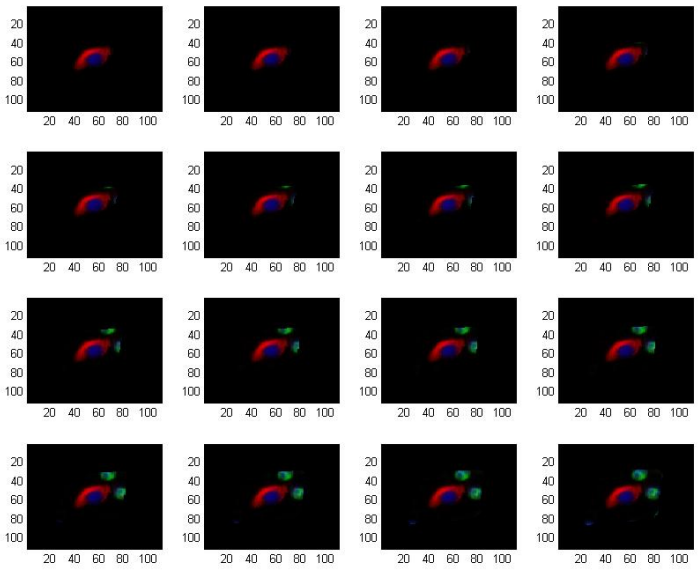


(a) (b) (c) (d) (e) (f)

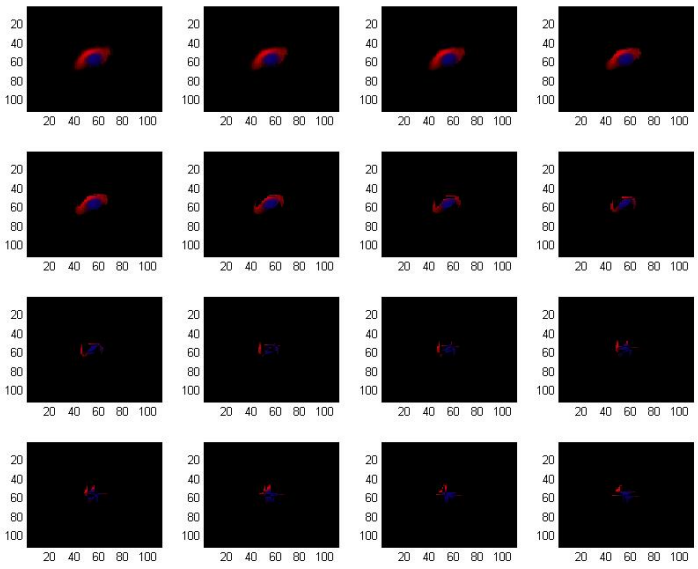
FIGURE 2.3. The set of images above show the lack of textural variation within the interior of the cells. The heat maps in (a)-(c) are from a sample HD-CTC while (d)-(f) come from a sample WBC. The lack of textural variation within the interior of a cell makes it very difficult to extract gradient based descriptors on the interior which reduces our ability to apply many extraction methods to this data set.

technician. We note that it would be considered acceptable to exceed the one hour time constraint if our method successfully identifies an event that was previously missed.

2.2.4. Segmentation. Segmentation issues are a reoccurring problem in cell classification tasks and several papers have been published on methods for improving segmentation in tasks of this variety. Two papers, in particular, discuss methods of segmentation for epithelial cells in blood [18, 19]. Often within the images we find that cells are overlapping or occur in clusters. Also, cells of low and high interest do not express themselves in the same image channels. Thus, there are many possible ways to segment events. We propose to design an algorithm which is robust to segmentation error on both the $DAPI$ channel (where current segmentation is performed) and on the entire cell. We test the relative robustness of each descriptor to simulated segmentation error by distorting the verticies of the manually drawn boundaries (as discussed in the following section) by various amounts and using the distorted boundaries as a new mask. The effects of segmentation are shown and discussed in the results section. The resultant images produced by distorting the boundary on a single cell, both shrinking and stretching by some number of pixels, are shown in Figure 2.4. The distortion of a boundary is done by taking the manually drawn boundary which consists of a set of verticies and either adding or subtracting some number of pixels from the coordinates of each vertex. To simulate a boundary which may contain neighboring cells we stretch the boundary by moving each vertex away from the center of the cell. In order to simulate a boundary which did not contain the entirety of the cell we shrink the boundary by moving each vertex in towards the center of the cell. The code for these distortions is provided in the Appendix.



(a)



(b)

FIGURE 2.4. (a) Shows the resultant curated images of a single HD-CTC as the manually drawn boundary is stretched by 1 to 15 pixels. (b) Shows the resultant curated images of the same HD-CTC as the manually drawn boundary is shrunk by 1 to 15 pixels.

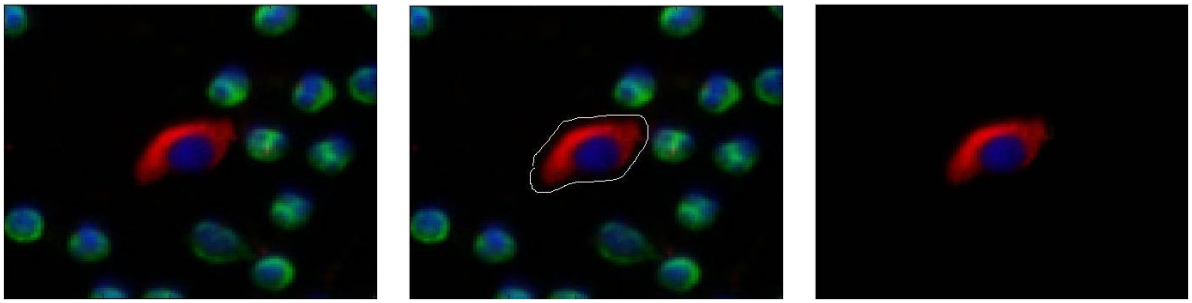
2.3. Curation of Data

Based on the inherent challenges of the data and results of preliminary classification attempts discussed in [9], we decided to follow a conservative approach and first test our methods on a manually curated data set. The data set consists of 1000 events. One half of the events had been previously labeled as events of interest by a technician at TKL while the other half are labeled as WBCs by us based on the attributes of a WBC as described to us by TKL.

Curated events of high interest were selected from four categories: HD-CTC, Apoptotic, No-CK, and Small-CK. Events were selected based on the quality of the image and the distance from neighboring events so as to avoid the previously mentioned issue of cell overlap. Additionally, there are no clusters of cells included in our manually curated data. There are 200 HD-CTC events and 100 events of each other high interest cell type.

Curated events of low interest were pulled from a total of 40 images. Two hundred events of low interest were selected in sets of 20 from 10 images that also contained an HD-CTC. The remaining 300 events of low interest were selected in sets of 10 from 30 images: 10 containing an Apoptotic event, 10 containing a No-CK event, and 10 containing a Small-CK event. These events were selected from relatively high distances away from the rare event contained in the image and were chosen to represent multiple visually discernible cell types to the eye with limited training.

Following the selection of the events a cell boundary was drawn by hand on the composite image. The composite image was selected to determine the boundary to ensure that the curated image contained all cell information which is most visible in the composite image. This cell boundary was then used to create a mask. True pixel values were kept from each



(a) (b) (c)

FIGURE 2.5. (a) Shows an HD-CTC in the original composite image surrounded by neighboring WBCs. (b) Shows the same HD-CTC with the manually drawn boundary, and (c) shows what a curated image of a single cell looks like after the boundary shown in (b) has been used to generate a mask and blackout all other events.

image channel if the pixel's location was within the boundary, otherwise the pixel value was set to zero. Last, the cell was centered within a 111x111 pixel cropped image. The size of the cropped image was selected to ensure full containment of all cells. Figure 2.5 shows the process of curating an image. The same boundary which was manually drawn was then used to test the effects of segmentation on the representations by distorting the boundary to generate a new mask and a new resultant curated image. The boundaries were both stretched and shrunk by up to 15 pixels as discussed and shown in the section discussing segmentation.

CHAPTER 3

Feature Extraction and Image Representations

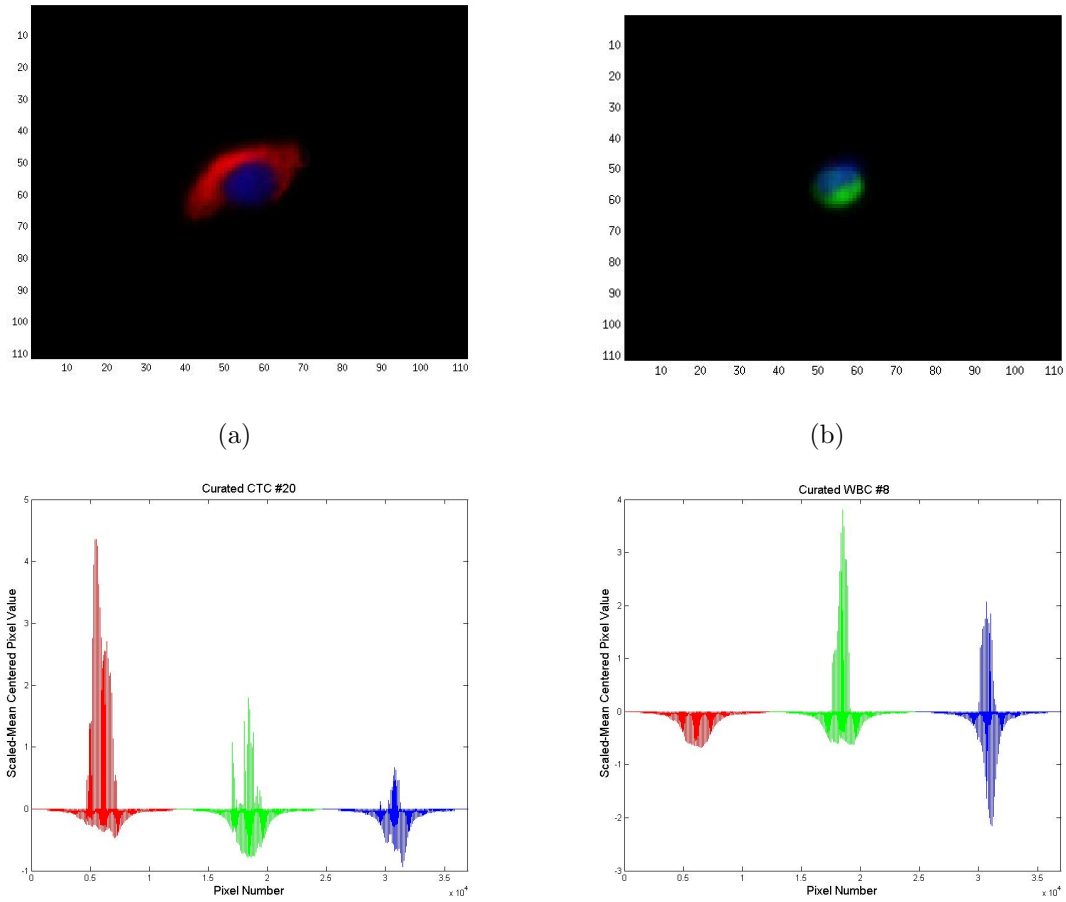
We have implemented several different feature extraction methods and image representation techniques including both codebook based descriptors and representations and non-codebook based methods. The discussion of this thesis will focus on non-codebook based techniques. Results and discussion of the codebook based implementations can be seen in [9]. In the following sections we will discuss a representation based on raw image values, a 2-dimensional Fourier transform descriptor, and a concentric ring based descriptor that exploits the rotational invariance of the Fourier transform, and finally the relationship of our methods to other work.

3.1. Raw Image Value

Both pathologists and the TKL algorithm utilize various measurements to determine a classification of an event as discussed in section 2.1.2. However, the measurements which contribute to a particular classification are strongly correlated to particular channel intensities. Since the intensities of the various channels are significant the first, rudimentary, representation of an event was to concatenate all pixel values. To this end we represent the i^{th} event, in a particular channel, as

$$\vec{v}^{(i)}_{channel} = vec(I^{(i)}_{channel})$$

where *vec* is the matlab command which concatenates all the columns of a matrix into a single column vector and $I^{(i)}_{channel}$ is the monochrome *channel* image of the i^{th} event. Thus, we have that $\vec{v}^{(i)}_{channel}$ is a 12321×1 vector representation of the event, recalling that each



(a) (b)

(c) (d)

FIGURE 3.1. (a) Shows a curated image of an HD-CTC while (b) shows a curated image of a WBC. (c) Shows the mean centered and scaled raw image value representation of the HD-CTC in (a). (d) Shows the mean centered and scaled raw image value representation of the WBC in (b).

image is 111×111 pixels in dimension and $111 \cdot 111 = 12321$, and where *channel* is chosen from $Alexa555$, $Alexa647$, or $DAPI$. The event is then represented as a 1×36963 vector defined as

$$\vec{V}_i = \begin{pmatrix} \vec{v}^{(i)}_{Alexa555} \\ \vec{v}^{(i)}_{Alexa647} \\ \vec{v}^{(i)}_{DAPI} \end{pmatrix}.$$

This image representation method has been utilized in other classification tasks with varying

degrees of success depending on the application. It is important to note that this image representation technique is heavily dependent on the quality of segmentation for the entire cell and is not rotationally invariant. Figure 3.1 show a visual representation of the raw image value representation of two sample cells: one of high interest and one of low.

3.2. 2-Dimensional Discrete Fourier Transform

The 2-dimensional discrete Fourier transform, implemented as the 2-dimensional fast Fourier transform (2D-FFT) in Matlab, has been implemented widely as a patch based feature descriptor. There are many beneficial properties of the 2D-FFT, including but not limited to, the ease of image reconstruction and the robustness of the descriptor to rotation.

A Fourier transform of an integrable function is a representation of a function f in terms of its frequency spectrum defined by

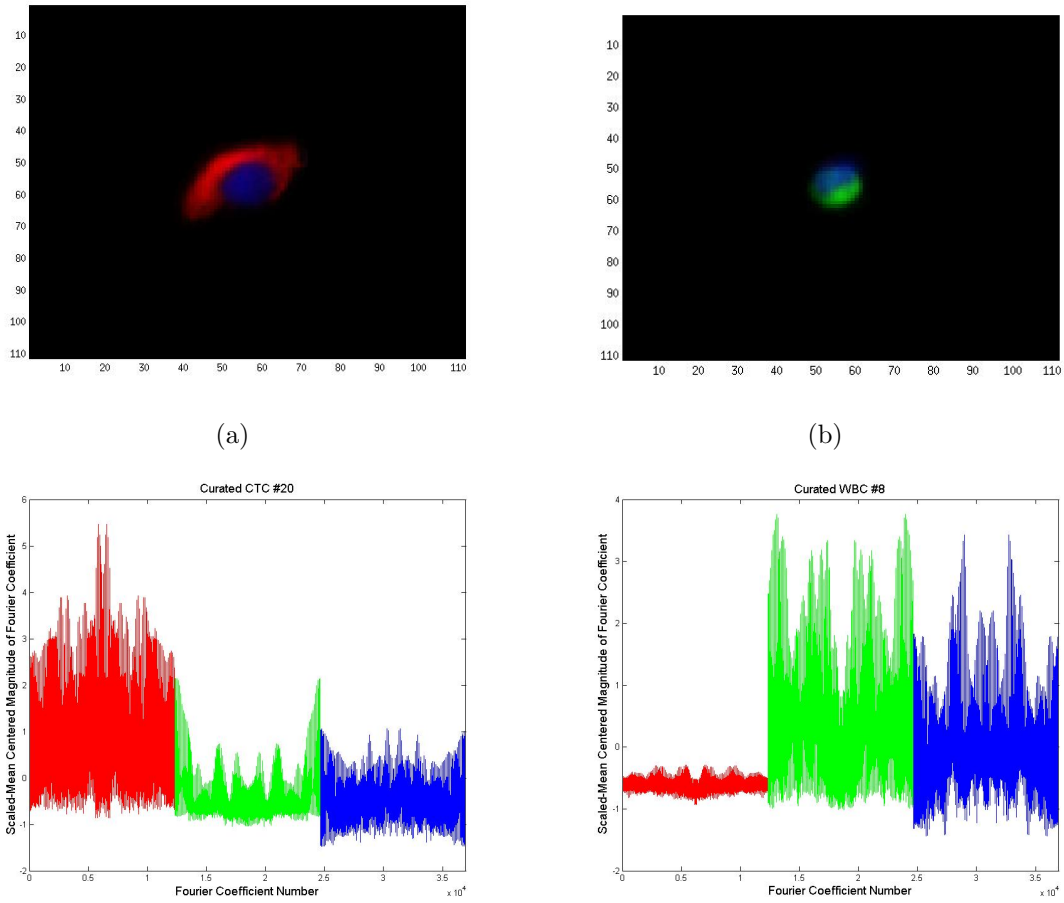
$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx$$

for ξ any real number. The Fourier transform represents a function as a combination of complex exponential functions, or equivalently as the combination of sine and cosine functions using Euler's formula

$$e^{i\theta} = \cos(\theta) + i \sin(\theta).$$

In the scenario where we wish to compute the Fourier transform of a discrete set of function values we can replace the integral with the Riemann sum. Thus, we use the following formula to compute the k^{th} term in the Fourier transform of $\vec{x}$, of length N, to be

$$X(k) = \sum_{j=1}^{N} x(j) \omega_N^{(j-1)(k-1)}$$



(a) (b)

(c) (d)

FIGURE 3.2. (a) Shows a curated image of an HD-CTC while (b) shows a curated image of a WBC. (c) Shows the mean centered and scaled 2D-FFT representation of the HD-CTC in (a). (d) Shows the mean centered and scaled 2D-FFT representation of the WBC in (b).

where

$$\omega_N = e^{-2\pi i/N}.$$

The k^{th} term of the Fourier transform is referred to as the k^{th} Fourier coefficient. Each Fourier coefficient is a complex number and for the sake of visualization we record the magnitude to be a real valued representation of the coefficient. It is sufficient to keep the magnitude of

the complex number since the distance between a pair of complex numbers and the distance between their magnitudes are equivalent.

When computing the 2D-FFT of an image of size $n \times m$ there will be $n \cdot m$ Fourier coefficients. We define the j, k 2D Fourier coefficient as

$$\hat{F}_{j',k'} = \frac{1}{MN} \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} F_{j,k} e^{-2\pi i(jj'/M + kk'/N)},$$

where $F_{j,k}$ is the image value of the j, k pixel. This is the formulation of the 2D-FFT as shown in *Geometric Data Analysis* [20]. To use 2D-FFT as an image representation of the i^{th} event we take the 2D-FFT as described above for each image channel and concatenate it into a vector. For each channel of a curated event we then obtain a 12321×1 vector

$$\vec{x}^{i}_{channel} = vec(\hat{I}^{(i)}_{channel}),$$

where $\hat{I}^{(i)}_{channel}$ is the 2D-FFT of the *channel* image of the i^{th} event, *channel* is chosen from $Alexa555$, $Alexa647$, or $DAPI$, and *vec* refers to the matlab command which stacks the columns of a matrix into a single vector. Next, we concatenate the transforms of each different channel to again obtain

$$\vec{X}_i = \begin{pmatrix} \vec{x}^{(i)}_{Alexa555} \\ \vec{x}^{(i)}_{Alexa647} \\ \vec{x}^{(i)}_{DAPI} \end{pmatrix}.$$

Figure 3.2 illustrates what the 2D-FFT representation of two sample cells look like. The implementation of this method to non-curated data requires segmentation of an event from the background and potentially some pre-processing to remove the noise created by directly

placing nonzero values in a true zero background. However, the generation of the representation does not require any previously computed values (i.e. center of mass of the cell, area, average image intensity, etc.).

3.3. Fourier Ring Descriptor

We propose a concentric ring based descriptor which also incorporates the fast Fourier transform and is therefore referred to as the Fourier Ring Descriptor (FRD). Unlike the 2D-FFT where a cell segmentation is required, the only previously computed value needed to begin the development of the FRD for an event is the center of mass of the cell nucleus. Given the center of a cell $\vec{c} = (x_{center}, y_{center})$, coordinates of points on a circle of radius r pixels centered at $\vec{c}$ are computed. The number of points on a circle of radius r is linearly scaled up from eight. Eight was chosen to be the number of points on a circle of radius $r = 1$ pixel because a circle of radius one will hit eight distinct pixels and therefore can correspond to at most eight distinct image values. Excessive sampling is redundant and computationally expensive. In our first implementation of this descriptor we used circles with radii varying from 1 to 25 pixels.

For the i^{th}event, the j^{th} point on a circle of radius r in a given channel we determine the image value at that location using bicubic interpolation and obtain the quantity $z_{i_{r_{j_{channel}}}}$. Thus, for each circle in a channel we obtain the vector

$$Z_{i_{r_{channel}}} = \begin{pmatrix} z_{i_{r_{1_{channel}}}} \\ z_{i_{r_{2_{channel}}}} \\ \vdots \\ z_{i_{r_{N_{channel}}}} \end{pmatrix},$$

where $N = 2 \times 8 \times r$ is the number of points on the ring of radius r. Next, for each $Z_{i_{r_{channel}}}$ the magnitudes of the coefficients of the fast Fourier transform of $Z_{i_{r_{channel}}}$ are computed as described in the previous section and stored as a vector $\hat{Z}_{i_{r_{channel}}}$. For each image channel we then concatenate the $\hat{Z}_{i_{r_{channel}}}$ in ascending order of the radii to obtain

$$\tilde{Z}_{i_{channel}} = \begin{pmatrix} \hat{Z}_{i_{1_{channel}}} \\ \hat{Z}_{i_{2_{channel}}} \\ \vdots \\ \hat{Z}_{i_{25_{channel}}} \end{pmatrix},$$

where channel is again chosen from $Alexa555$, $Alexa647$, or $DAPI$. Lastly, we concatenate the concatenated magnitudes of the Fourier coefficients by channel to obtain a final image representation of the i^{th} event to be

$$\mathring{Z}_i = \begin{pmatrix} \tilde{Z}_{i_{Alexa555}} \\ \tilde{Z}_{i_{Alexa647}} \\ \tilde{Z}_{i_{DAPI}} \end{pmatrix}.$$

After preliminary work using 25 concentric rings, we observed that there were no features being selected for the classifier that came from outside the sixteenth ring in the $Alexa555$ and $DAPI$ channels and no features selected outside the twelfth ring in $Alexa647$. This is why the Figures 3.4 and 3.5 show sixteen rings. The number of rings required corresponds to circles of smaller radii than the TKL algorithm uses. In the case of the FRD descriptor there is minimal reliance on the segmentation of the entire cell and it is therefore more readily implementable to non-curated data. Contrary to the other two descriptors, this descriptor

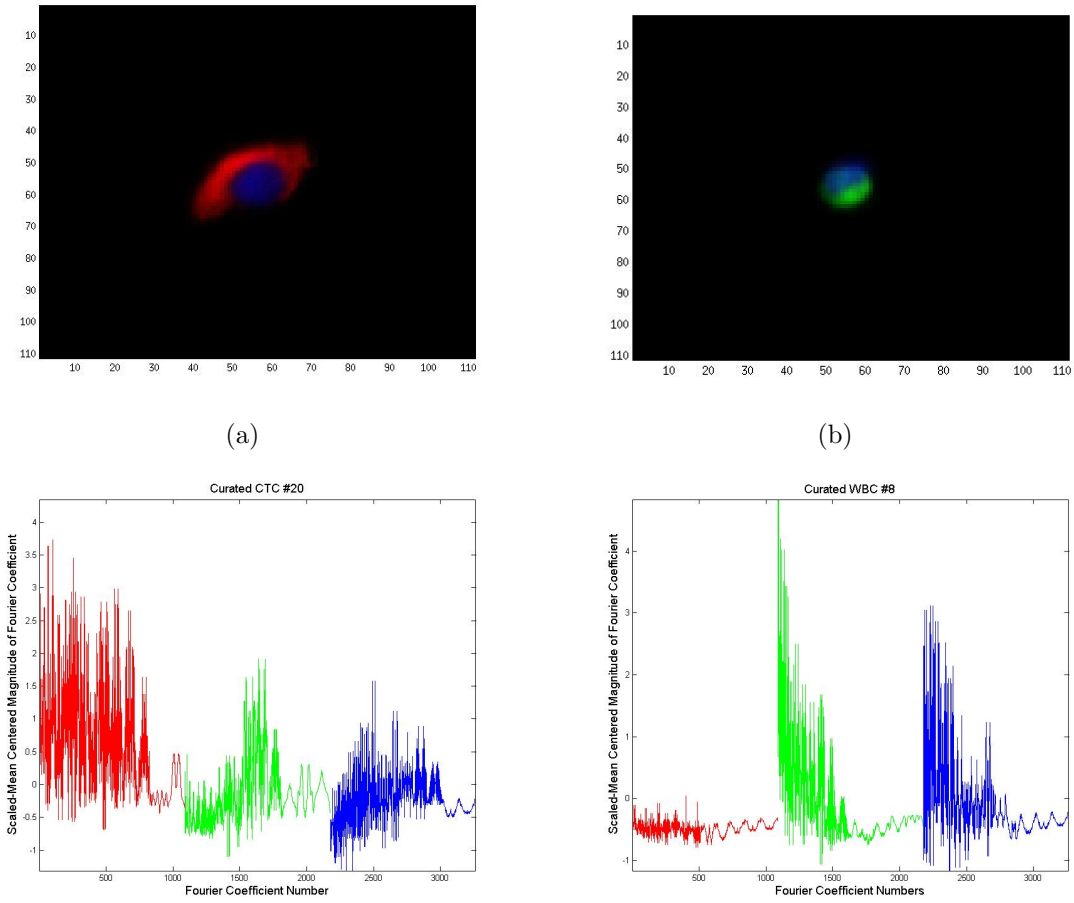
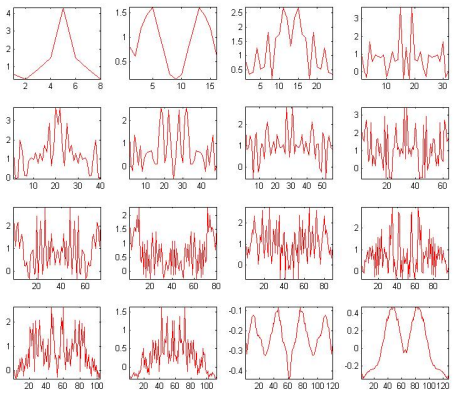
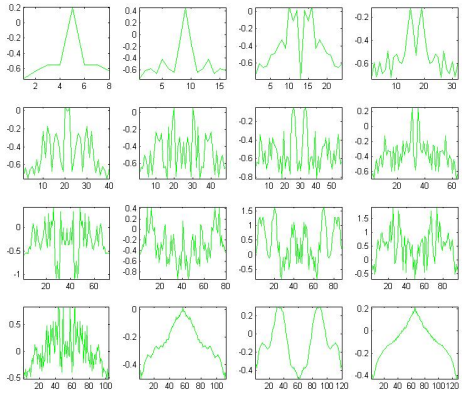


FIGURE 3.3. (a) Shows a curated image of an HD-CTC while (b) shows a curated image of a WBC. (c) Shows the mean centered and scaled FRD representation of the HD-CTC in (a). (d) Shows the mean centered and scaled FRD representation of the WBC in (b).

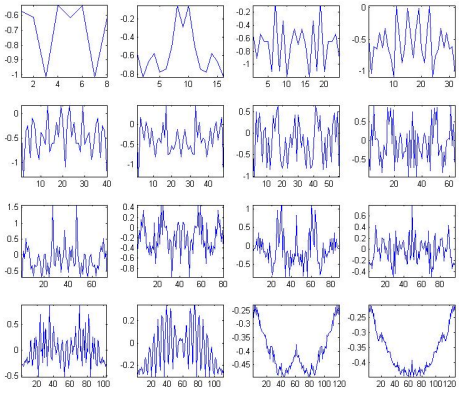
does require the computation of center of mass of the cell. However, this computation is done once on a single channel per event and is thus readily implementable. Since we encountered issues of rotational invariance in [9] we wanted to ensure that our proposed descriptor was rotationally invariant. A function which is periodic is invariant to translation, rotation, and shifting. As stated in **Proposition 5.2** in [20], "The amplitude spectrum is periodic." Thus, since we are writing the images values along a circle in terms of their amplitude spectrum we



(a)



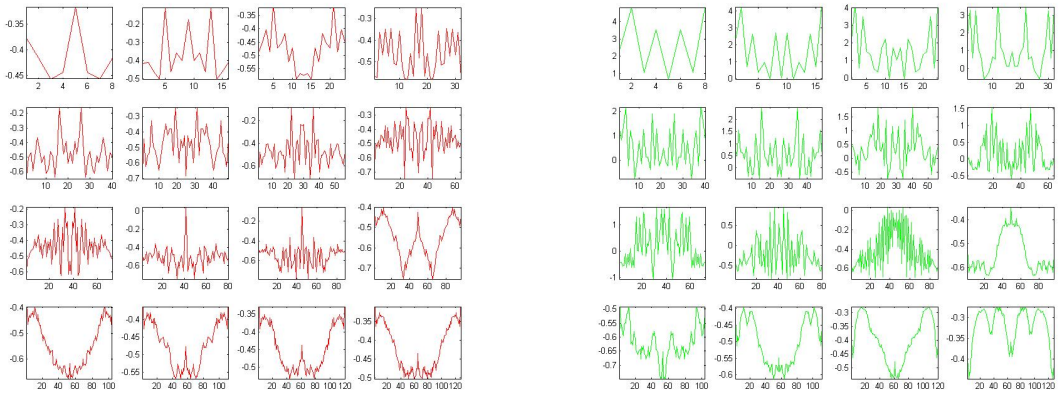
(b)



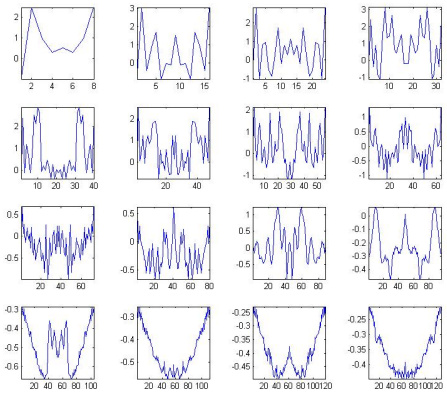
(c)

FIGURE 3.4. Shows the magnitudes of the fast Fourier transform of each individual of the 16 concentric rings in the different channels for the HD-CTC shown in Figure 3.3. The $x-axis$ in each plot is the Fourier coefficient number in a particular ring and the $y-axis$ is the mean centered and scaled magnitude of the Fourier coefficient. The subplots correspond to the various radii used starting with 1 in the upper left corner, increasing left to right and top to bottom. (a) Corresponds to the rings extracted from the $Alexa555$ channel, (b) corresponds to the $Alexa647$ channel, and (c) corresponds to the $DAPI$ channel.

have that $\tilde{Z}_{i_{channel}}$ is rotationally invariant. Code for the computations shown can be found in the appendix.



(a) (b)



(c)

FIGURE 3.5. Shows the magnitudes of the fast Fourier transform of each individual of the 16 concentric rings in the different channels for the WBC shown in Figure 3.3. The $x - axis$ in each plot is the Fourier coefficient number in a particular ring and the $y - axis$ is the mean centered and scaled magnitude of the Fourier coefficient. The subplots correspond to the various radii used starting with 1 in the upper left corner, increasing left to right and top to bottom. (a) Corresponds to the rings extracted from the $Alexa555$ channel, (b) corresponds to the $Alexa647$ channel, and (c) corresponds to the $DAPI$ channel.

3.4. Relation to Other Work

We were motivated to construct a Fourier based feature descriptor based on properties intrinsic to our data and to have a descriptor that was rotationally invariant. Other

approaches to this problem have been taken. For example spatially invariant vector quantization (SIVQ) is a pattern matching algorithm developed by Dr. Jason Hipp. The SIVQ descriptor methods advertises rotational invariance and states that concentric rings act like a safe combination lock [21]. Hipp expands on this metaphor by explaining that each ring can be rotated independently of the others. In allowing this possibility the method requires a codebook like structure where there is a set of "code" rings corresponding to each of the radii used. What is inferred from reading [21] is that there is an exhaustive search performed when identifying the closest code ring for a ring from a novel sample. This is done by first breaking the circular descriptor, which appears to be a set of interpolated image values along the ring, at an arbitrary point and finding the nearest code ring. Next, a shift of the circular descriptor is performed and the nearest code ring is again found. The method continues until all circular permutations have been tried and the ring is assigned to the code ring corresponding to the minimal distance over all permutations. This is then done for each ring. After each ring has been matched to its code ring, the information is put into a vector representation and classified.

The SIVQ pattern matching technique has had success in several applications [21], but ultimately did not seem applicable to our classification task for several reasons. First, in the applications where SIVQ showed great promise there was a significant amount of textural variation in the images which is not present in our data. Secondly, SIVQ was largely successful in identifying events once a small sample had been obtained; SIVQ was not used to pull out candidates of interest from large data sets. Additionally, the exhaustive nature of ring matching described in the implementation of SIVQ is infeasible at the scale of the number of events needing to be classified in the problem we describe.

While SIVQ is rotationally invariant it is computationally expensive. We therefore wanted to incorporate the concentric ring idea with the Fourier transform which is also rotationally invariant and less costly than exhaustive permutations to find a best match. Discrete Fourier transforms can be used in many ways to aid in classification of cells. We have discussed two ways we have employed a discrete Fourier transform (2D-FFT and FRD), but there are other ways to utilize Fourier transforms. For example, a recent paper on the automated classification of images of human epithelial cells used 2D-FFT as a patch based feature extractor as a component in a codebook image representation technique [22]. In [22] they seek to classify an entire image of cells as a single class and therefore the extraction of a small number of patches on each cell within the image is not a cause for large concern. However, based on the limited size of the events of interest in our study and that we seek to classify a single cell within an image, we implement the 2D-FFT as a complete image representation instead of as a patch based method.

A ring based descriptor exploits the differences in morphology of events of low and high interest without having to compute measurements of circularity. Using concentric rings of varying size also allows us to exploit the natural size variation of cells without having to do area computations. Also, when implemented as we describe our ring based descriptor is rotationally invariant and does not require codebooks or exhaustive comparison.

CHAPTER 4

Classification

A primary focus of our research is to develop a series of decision functions to determine the class of particular cell. The first set of classes are known cell types, while later we hope to determine additional classes that are not currently labeled. Given a cell we first want to decide whether it is an event of high interest(EOHI) or if it is a white blood cell (WBC). For this classification task we refer to an EOHI as a member of C^+ and a WBC as a member of C^-. After a cell has received its classification as a member of either C^+ or C^- we then want to determine what subclass the cell belongs to. We have four known subclasses of the

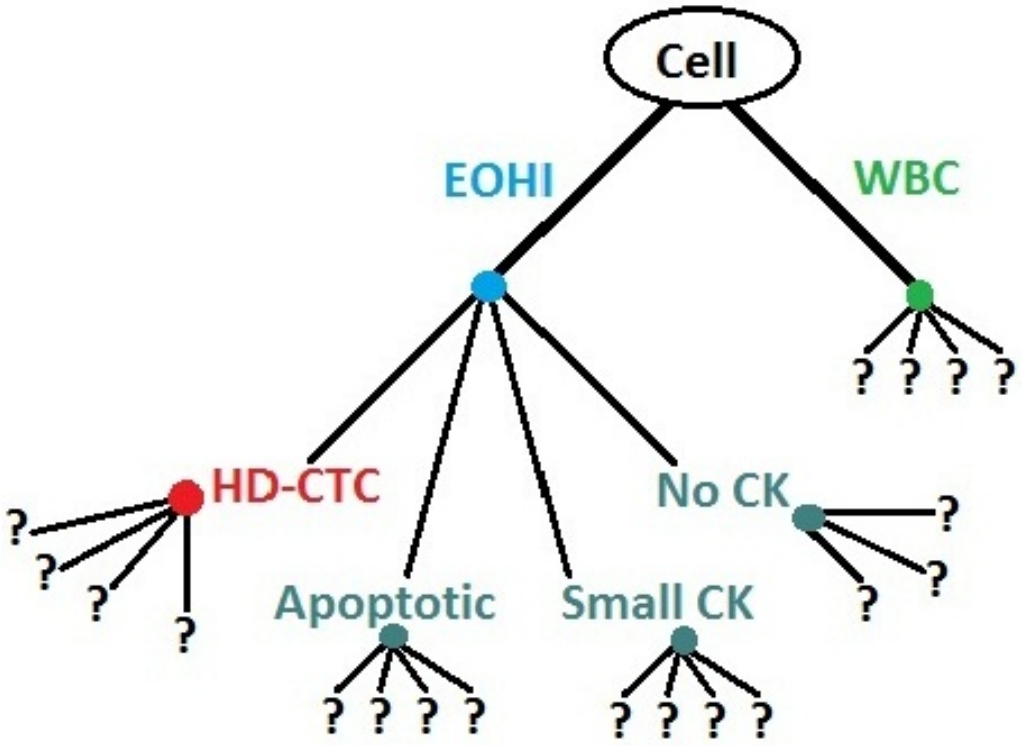


Figure 4.1. This figure provides an illustration of the structure of the decision tree we are trying to develop for the the classification of cells. The number of branches ending in question marks off of each known subset is illustrative only.

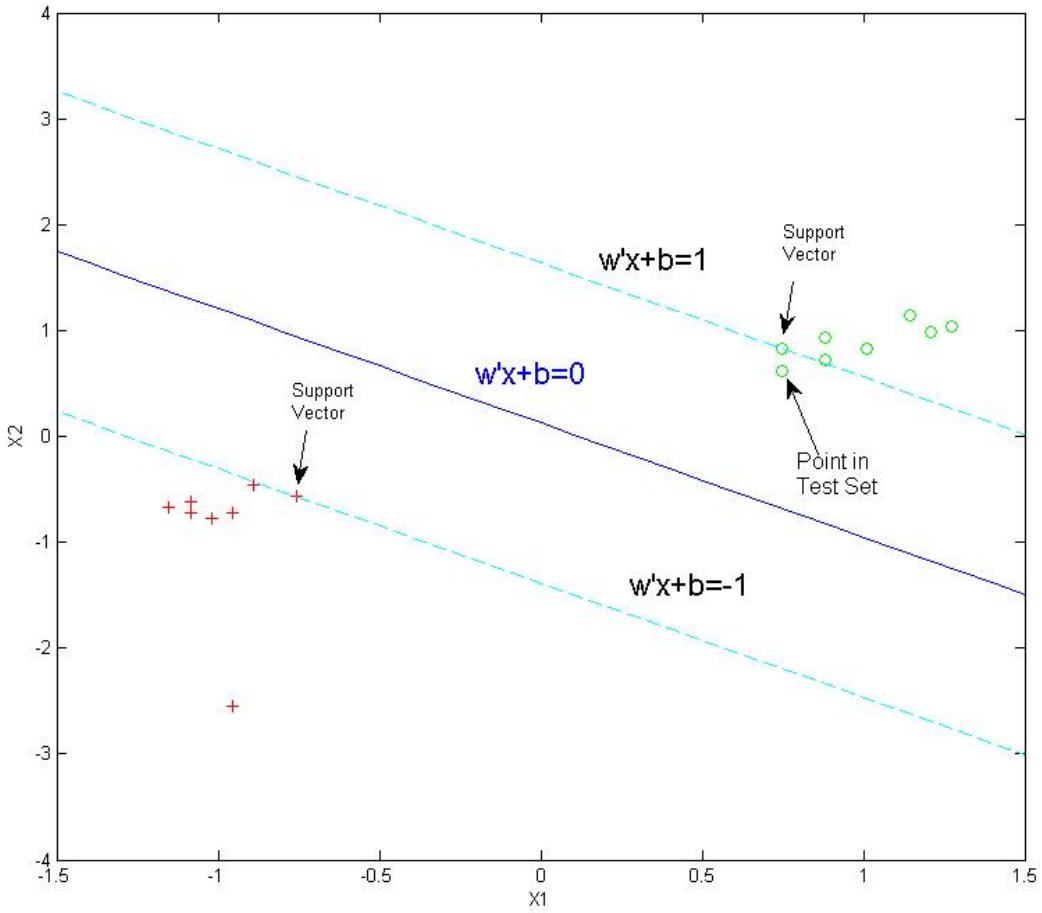


FIGURE 4.2. This figure provides an illustration of the hyperplane separating two classes in two dimensions. These classes are linearly separable and have no events misclassified.

class C^+ and currently no known subclasses of C^-. In previous work done on this problem, discussed in [9], we implemented hierarchal clustering algorithms for our decision functions to discern unknown subclasses and essentially began at the bottom of the decision tree model shown in Figure 4.1. These methods are computationally expensive. While we hope to continue exploring additional types of decision functions, especially when we return to the unsupervised problem of determining unknown subclassifications of cells, we have made use

of readily available tools for the supervised problems corresponding to labeled branches of Figure 4.1.

For the two class classification problem of events of high versus low interest, on both curated and non-curated data, we have used the open source support vector machine (SVM) made available by LibLINEAR [23] with a linear kernel, l_1 regularized classifiers and an l_2 loss function. In doing so we aim to determine a hyperplane which linearly separates our data classes while maximizing the distance to the nearest points of each class from the hyperplane symmetrically, as shown in Figure 4.2, by solving the optimization problem

$$\min_{w} \|w\|_1 + C \sum_{i=1}^{n} max(0, 1 - y_i w^T x_i)^2.$$

In the optimization problem above w is the normal vector to the hyperplane, $\{(x_i, y_i)\}_{i=1}^{n}$ is the set of data points, and C is a penalty parameter. Additionally, in the LibLINEAR implementation the cost function is a combination of the one norm of the normal vector and a two norm lost function depending on whether you want to use the one norm or two norm loss function.

The points closest to the separating hyperplane are called the support vectors. The vector $\vec{w}$ in Figure 4.2 is the normal vector to the plane. An event $\vec{x}$, not in the set used to build the classifier, is classified by the sign of $\vec{w}^T\vec{x} + b$. In Figure 4.2 the classes are linearly separable and no events are misclassified. This is a very simple example that does not cover many of the possible problems that can arise. Our current implementations have weighted each class equally but in the future we will explore making the cost of EOHI misclassification more expensive.

We have chosen to use an l_1 regularized classifier in order to force sparsity of features. This is important in the case of the 2D-FFT and RIV descriptors which contain $36,963$

features. When sparsity is forced it often means that the set of features selected to obtain high accuracy is not unique. A further discussion of this follows in the results section. Additionally, there are arguments for and against mean centering and scaling data before performing SVM classification. We have applied a linear SVM classifier to unaltered data and to mean-centered and scaled data. A discussion of the effect of mean-centering is included in the results section.

The l_1 regularized, l_2 loss function linear SVM can be solved by using primal-dual interior point methods. Given the primal form of the problem

$$\min_w \|w\|_1 + C\sum_{i=1}^{n} max(0, 1 - y_i w^T x_i)^2.$$

we can introduce additional variables which allow us to rewrite the primal form in such a way that it can be easily differentiated. First, we can let $w = w^+ - w^-$. We can then rewrite $\|w\|_1 = e^T w^+ + e^T w^-$ where e is the vector of all ones of the appropriate length. Additionally, we can define $\xi_i = 1 - y_i w^T x_i$, or equivalently, $\xi_i = 1 - y_i(w^+ - w^-)^T x_i$, where the y_i is the class label of the data point x_i. Lastly, we can define a diagonal matrix D with the y along the diagonal, and a matrix X with the data points x_i as rows in the matrix. By making these changes the primal problem becomes

$$\text{minimize } e^T w^+ + e^T w^- + c\xi^T\xi$$

$$\text{subject to } DX(w^+ - w^-) \geq e - \xi$$

$$w^+, w^-, \xi \geq 0$$

With this primal problem we can then compute the Lagrangian function L as

$$L = e^T w^+ + e^T w^- + c\xi^T - \alpha^T(DX(w^+ - w^-) - e + \xi) - \beta^T w^+ - \gamma^T w^- - \mu^T \xi.$$

Once we have the Lagrangian we can derive the Karush-Kuhn-Tucker (KKT) Conditions based on the function being stationary, requiring both primal and dual feasibility, and looking

at complementary slackness conditions to hold. The function being stationary refers to all the partial derivatives of the Lagrangian to be equal to zero. From this component of the KKT conditions we obtain the following conditions:

$$\frac{\delta L}{\delta w^+} = e - X^T D\alpha - \beta = 0 \Rightarrow \beta = e - X^T D\alpha$$

$$\frac{\delta L}{\delta w^-} = e + X^T D\alpha - \gamma = 0 \Rightarrow \gamma = e + X^T D\alpha$$

$$\frac{\delta L}{\delta \xi} = 2c\xi - \alpha - \mu = 0 \Rightarrow \mu = 2c\xi - \alpha$$

For primal feasibility we require

$$DX(w^+ - w^-) - e + \xi \geq 0$$

$$w^+, w^-, \xi \geq 0.$$

From dual feasibility we simply require $\alpha, \beta, \gamma, \mu \geq 0$ which yields the following constraints

$$\beta, \gamma \geq 0 \Rightarrow -e \leq X^T D\alpha \leq e$$

$$\mu \geq 0 \Rightarrow 2c\xi - \alpha \geq 0$$

$$\alpha \geq 0 \Rightarrow 0 \leq a \leq 2c\xi$$

Lastly, we look at the conditions enforced by complementary slackness. Complementary slackness requires that $\alpha^T(DX(w^+ - w^-) - e + \xi) = 0$, $\beta^T w^+ = 0$, $\gamma^T w^- = 0$, and that $\mu^T \xi = 0$. Plugging all of the relationships that are derived from the KKT conditions into the original Lagrangian and simplifying will then allow us to obtain the formulation of the dual problem. Once we have the paired Primal-Dual problems, we can apply one of the Primal-Dual interior methods to reach an optimal solution, if one exists.

CHAPTER 5

Results

In this chapter we present the results of three methods applied to two different tasks. First we will compare the performance of each of the three proposed descriptors on the two class problem separating events of high interest (EOHI) from white blood cells (WBCs). Next, we will discuss the performance of each descriptor on the pairwise classifications tasks of subdividing EOHI into four subclasses. In the following sections the expression "mean-centered" refers to the matrix consisting of all descriptors of one type being column wise mean centered and scaled by the standard deviation, while "unaltered" refers to the matrix consisting of all descriptors of one type without any interference. All explorations were performed on both mean-centered and unaltered versions of the descriptors. This was done largely to help determine whether the variation of stain expression is important to the classification tasks at hand. Also, in order to fully interpret the results shown it is important to note that when the descriptors are mean centered the number of features required for classification is reduced, often by a factor between 2 and 10.

5.1. Comparison on the EOHI vs. WBC two Class Problem

The first set of results shown in Tables 5.1 and 5.2 show a comparison of the three image representations in terms of their performance averaged over 500 randomly generated trials on the two class problem where we seek to separate all EOHI from WBCs. In each trial 75% of the data of each class is randomly selected and used for trainiing (375 events), while the remaining 25% is used for testing (125 events). The 'Total Number of Distinct Events Ever

TABLE 5.1. Results of Representations on Two Class Problem (Unaltered)

	FRD	2D-FFT	RIV
Average Overall Accuracy	$99.3800 \pm 0.4337\%$	$99.4272 \pm 0.5827\%$	$99.2080 \pm 0.5072\%$
Average Number of EOHI Misclassified as WBC	1.0100 ± 1.0049	0.6840 ± 0.8174	0.9400 ± 0.9780
Average Number of WBC Misclassified as EOHI	0.5400 ± 0.6726	0.7480 ± 0.7619	1.0400 ± 0.8800
Total Number of Distinct Events Ever Misclassified	19	21	45
Number of Distinct HD-CTC Ever Misclassified	2	5	2
Number of Distinct Apoptotic Ever Misclassified	1	1	7
Number of No CK Ever Misclassified	8	8	25
Number of Small CK Ever Misclassified	0	1	3
Maximum Number of EOHI Misclassified in Single Trial	4	4	6
Average Time to Generate Classifier (in Seconds)	0.6478	6.8641	0.8117

TABLE 5.2. Results of Representations on Two Class Problem (Mean-Centered)

	FRD	2D-FFT	RIV
Average Overall Accuracy	$99.3416 \pm 0.4548\%$	$99.2664 \pm 0.5100\%$	$99.4360 \pm 0.3976\%$
Average Number of EOHI Misclassified as WBC	0.9140 ± 0.9101	1.1620 ± 1.1019	0.4800 ± 0.6530
Average Number of WBC Misclassified as EOHI	0.7320 ± 0.7858	0.6700 ± 0.7782	0.9300 ± 0.8641
Total Number of Distinct Events Ever Misclassified	31	49	24
Number of Distinct HD-CTC Ever Misclassified	8	12	3
Number of Distinct Apoptotic Ever Misclassified	2	1	5
Number of No CK Ever Misclassified	7	14	6
Number of Small CK Ever Misclassified	3	8	2
Maximum Number of EOHI Misclassified in Single Trial	4	5	3
Average Time to Generate Classifier (in Seconds)	0.6484	6.2560	3.5212

Misclassified' refers to the number of events that were misclassified in one or more of the 500 trials.

In the Table 5.1, the average accuracy of each of the representations are not statistically significantly different. Nor are their relative numbers of false positive and false negatives. However, the total number of events misclassified over 500 trials makes the Fourier based methods stand out as they give false classifications less than half the number of events as RIV over 500 trials. This measure illustrates the robustness of the models generated by each image representation. Thus, Table 5.1 suggests that FRD generates the most robust models. Although we want to correctly classify all EOHI, our top priority is to correctly identify HD-CTCs. To this end FRD and RIV are the representations that minimize the number of misclassified HD-CTCs over 500 trials and achieve our top priority. Additionally, if one seeks to select the fastest classifier for this classification task, FRD wins by a factor of roughly ten over 2D-FFT on both versions of the descriptors and a factor of about 6 over RIV on mean-centered descriptors and a factor of about 1.5 over RIV on unaltered descriptors. This is an important measurement to consider when we think of the large scale implementation.

Results shown in Table 5.2 were generated under the same conditions as Table 5.1 with the exception that the descriptors were mean centered prior to classification. While the average accuracy, false positives, and false negatives of each model does not change in a statistically significant way between mean-centered and unaltered descriptors, the total number of events misclassified roughly doubles in the Fourier representations and halves in RIV. Also, in all three representations the number of HD-CTCs misclassified goes up. Furthermore, an effect of mean centering is an increase in the time taken to generate a classifier for RIV. While it is often considered a standard technique to mean-center and scale data prior to using a

TABLE 5.3. Comparison of Average Number of Selected Features

	Mean-Centered and Scaled	Unaltered
FRD	227	405
2D-FFT	541	1757
RIV	228	325

SVM, our results suggest that on this particular problem the variation of intensities between cells in different images contains information useful to the EOHI versus WBC classification task. As a point of curiosity we generated the FRD descriptors for the cells in our curated data set from their uncurated form and generated 500 classifiers randomly partitioning the data. When FRD is used on uncurated images of hand selected events, it averages a correct classification of $99.5328 \pm 0.4199\%$ on unaltered descriptors and $99.5656 \pm 0.4181\%$ on mean-centered and scaled descriptors. To obtain a comprehensive view of the performance of each representation we must also look at the robustness of the features selected for classification. As previously mentioned, using an l_1 classifier forces sparsity of features used in classification. One can then inquire as to whether or not the selected set of features for classification is unique. The Figures 5.1 and 5.2 show the average effects of removal of feature sets on the accuracy of a model in the EOHI vs. WBC problem over 500 trials. Also as previously stated, when the descriptors are mean centered the number of features required for classification goes down. The average number of features selected, when the classifier is based on all possible features, for each trial on both mean centered and unaltered data are shown in Table 5.3.

For example, the average number of features, selected to generate a classifier for FRD drops from around 400 on unaltered descriptors to around 220 on mean-centered and scaled descriptors. This is why the x-axis of Figure 5.1 extends only to 10 instead of 50. At the lowest end, we would consider a 95% classification to be acceptable. The average accuracy of the models generated using unaltered descriptors falls below this threshold in all cases

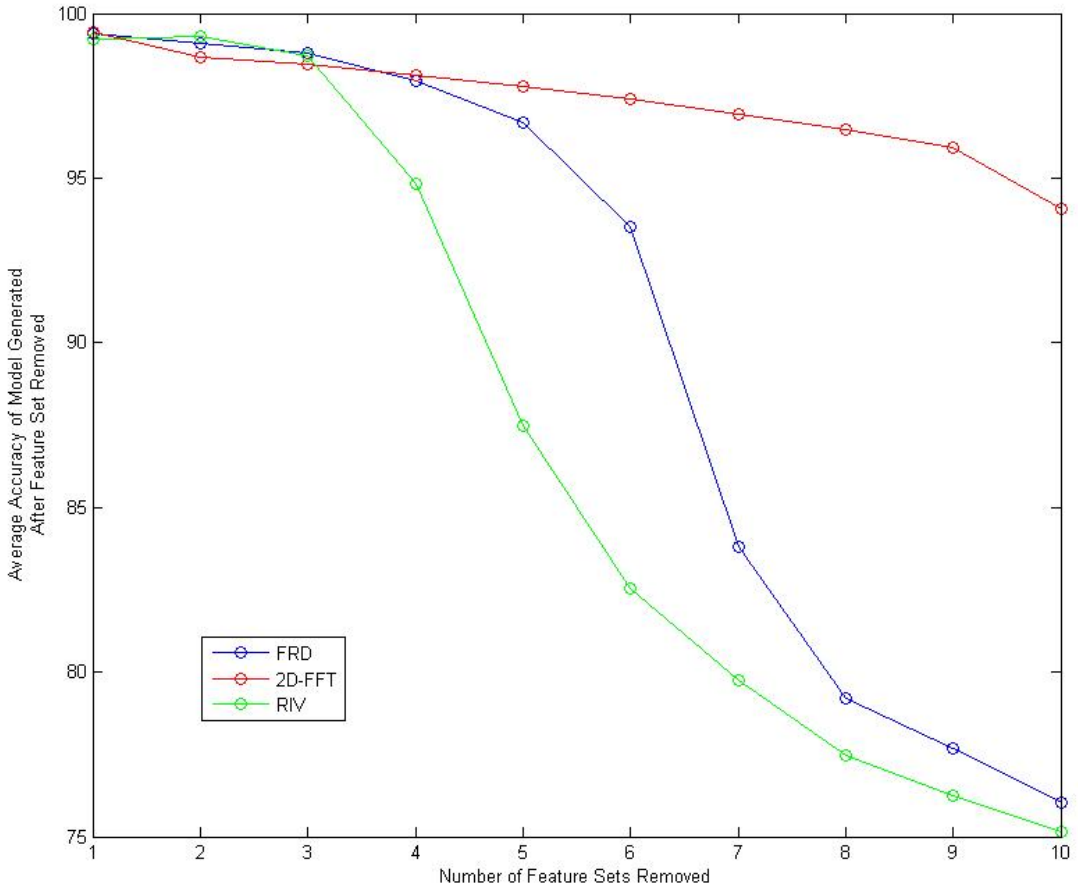


FIGURE 5.1. The average accuracy over 500 randomly generated models as the first ten selected feature sets are removed on unaltered descriptors.

after the first ten selected feature sets have been removed. On the mean-centered and scaled descriptors, however, the average accuracy of the 2D-FFT descriptor maintains an acceptable level even after the first 50 selected feature sets have been removed. It is also important when interpreting these results to recall that FRD has 3264 total features while 2D-FFT and RIV have 36963 features before selection. This is important, for example, in the unaltered FRD case where if 300 features were selected at each step, after five steps 50% of all the features would have been removed. If the same number of features were selected in each step in 2D-FFT or RIV, after five steps only about 5% of all features would have been removed.

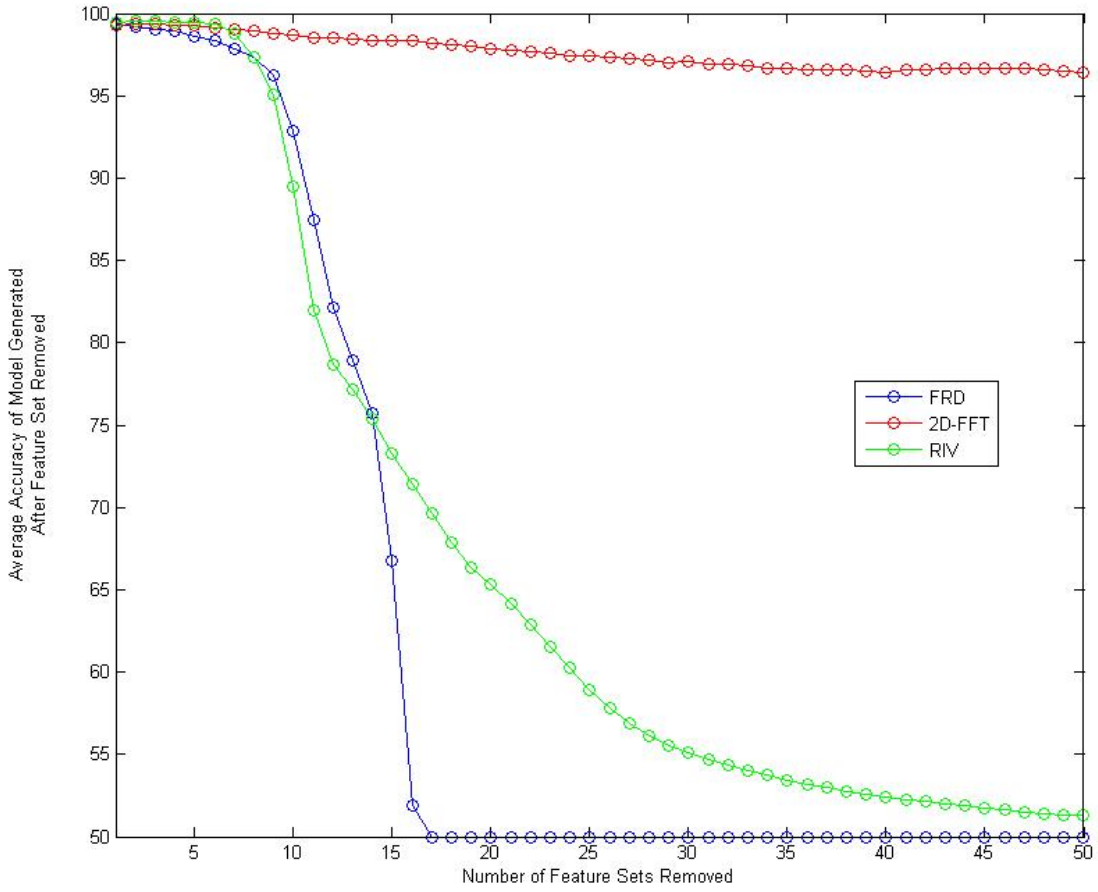
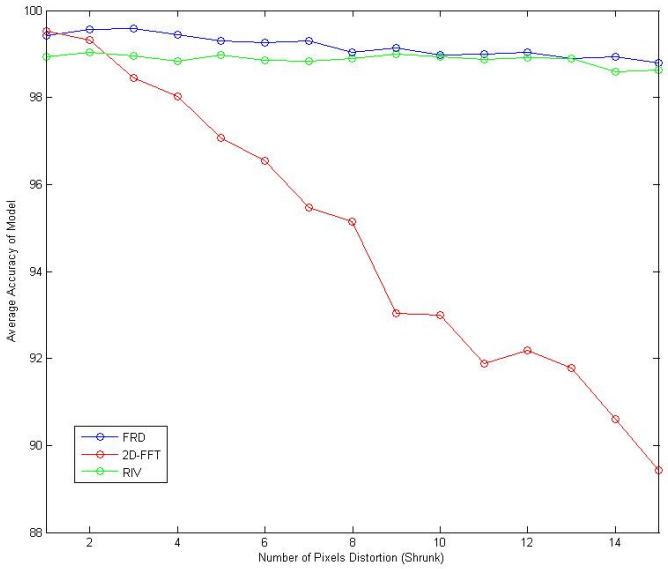
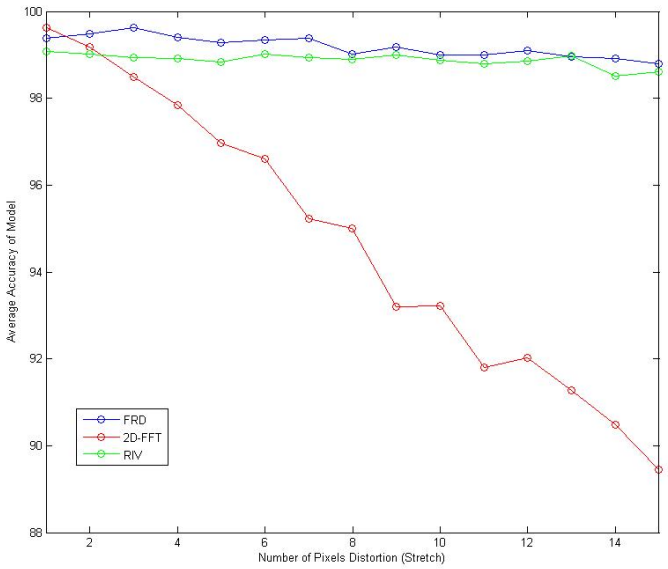


FIGURE 5.2. The average accuracy over 500 randomly generated models as the first fifty selected feature sets are removed on scaled and mean-centered descriptors.

Another component of the performance of the various representations is the effects of different levels of segmentation error on each of the three methods. The effects of segmentation on all methods on both mean-centered and scaled data as well as unaltered data are shown in Figures 5.3 and 5.4, respectively. Whether the boundary is shrunk or stretched, FRD maintains the highest average accuracy in almost all cases when mean-centered and scaled or unaltered. The performance of 2D-FFT falls off the most rapidly in both cases. This is only one way of testing the effects of segmentation on the methods, and while it shows that

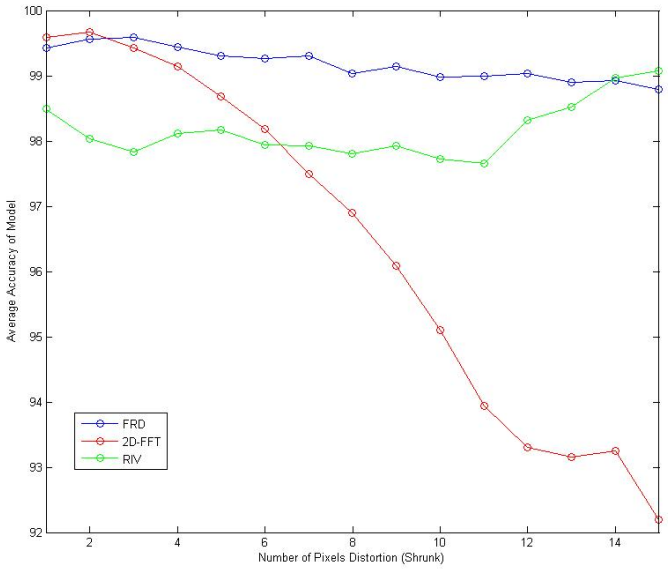


(a)

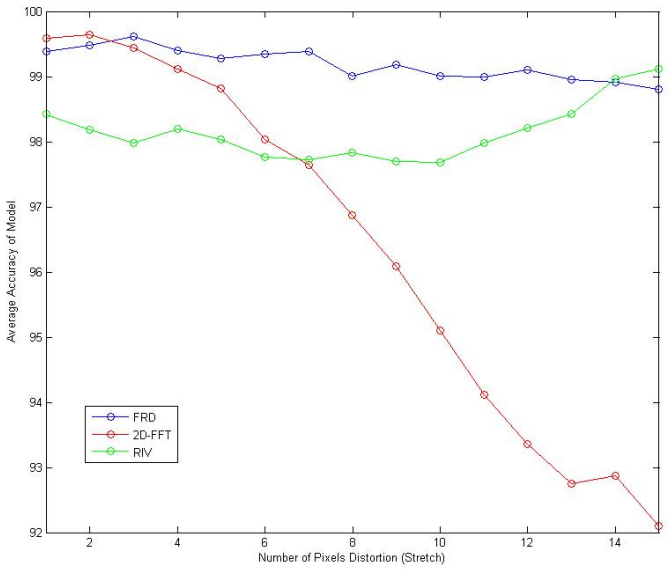


(b)

FIGURE 5.3. Comparing the average accuracy over 100 randomly generated models for different levels of distortion of the masking boundary for each of the three representations. These models were generated on unaltered descriptors. (a) Shows a comparison of the three methods as the masking boundaries is shrunk. When a boundary is shrunk it represents segmentation error where the CK and CD-45 information are reduced. (b) Compares the three representations as the masking boundaries are stretched which corresponds to segmentation error as neighboring cells are included in the boundary.



(a)



(b)

FIGURE 5.4. Comparing the average accuracy over 100 randomly generated models for different levels of distortion of the masking boundary for each of the three representations. These models were generated on mean-centered and scaled descriptors. (a) Shows a comparison of the three methods as the masking boundaries is shrunk. When a boundary is shrunk it represents segmentation error where the CK and CD-45 information are reduced. (b) Compares the three representations as the masking boundaries are stretched which corresponds to segmentation error as neighboring cells are included in the boundary.

the FRD descriptor is the most robust to this type of segmentation further analysis would be necessary before broad scale application to other cell tasks.

5.2. Pairwise EOHI Classification

There are six different pairwise EOHI classifications we can discuss:

- HD-CTC vs. Apoptotic,
- HD-CTC vs. No CK,
- HD-CTC vs. Small CK,
- Apoptotic vs. No CK,
- Apoptotic vs. Small CK,
- Small CK vs. No CK.

In this section we discuss the performance of each of the three descriptors on both mean-centered and unaltered data on each of the six pairwise classification tasks.

A comparison of the average performance over 100 randomly generated trails of each descriptor on the pairwise EOHI classification problems, where 75% of each class is used for training and 25% for testing, on both mean-centered and scaled descriptors as well as unaltered descriptors are shown in Tables 5.4 and 5.5 respectively. Again, we see no truly statistically significant differences between the performance of the methods on each pairwise classification task in most cases, or between a single method's performance when the descriptors are mean-centered and scaled or unaltered, but we note that the FRD descriptor has the highest average performance and usually the smallest standard deviation for each of the six pairwise classification tasks.

Each of the three near HD-CTC populations are considered marginal populations for specific reasons. For example, No CK is a population of cells that has the appropriate

TABLE 5.4. Results of Representations on Pairwise Classification of EOHI (Mean Centered)

	FRD	2D-FFT	RIV
HD-CTC vs. Apoptotic	$93.6200 \pm 3.4547\%$	$84.3400 \pm 5.2498\%$	$86.3200 \pm 4.6772\%$
HD-CTC vs. No CK	$89.5600 \pm 3.9449\%$	$87.8200 \pm 4.6718\%$	$88.2000 \pm 4.4812\%$
HD-CTC vs. Small CK	$94.0200 \pm 3.5219\%$	$89.9200 \pm 4.2554\%$	$91.4600 \pm 3.7266\%$
Apoptotic vs. No CK	$90.7800 \pm 4.1181\%$	$87.3800 \pm 3.8841\%$	$81.3400 \pm 6.0507\%$
Apoptotic vs. Small CK	$82.4400 \pm 4.7019\%$	$81.9800 \pm 5.4938\%$	$77.0200 \pm 4.9113\%$
Small CK vs. No CK	$96.5600 \pm 2.3455\%$	$96.2200 \pm 2.7692\%$	$96.9400 \pm 2.0391\%$

TABLE 5.5. Results of Representations on Pairwise Classification of EOHI (Unaltered)

	FRD	2D-FFT	RIV
HD-CTC vs. Apoptotic	$95.4200 \pm 2.5153\%$	$89.4400 \pm 3.7721\%$	$87.6800 \pm 4.9785\%$
HD-CTC vs. No CK	$93.3000 \pm 3.0134\%$	$91.9800 \pm 3.7023\%$	$90.6800 \pm 3.6317\%$
HD-CTC vs. Small CK	$95.0800 \pm 2.7768\%$	$93.4600 \pm 3.2518\%$	$93.4800 \pm 3.4421\%$
Apoptotic vs. No CK	$92.6400 \pm 3.3136\%$	$86.5600 \pm 3.8985\%$	$75.8000 \pm 5.4643\%$
Apoptotic vs. Small CK	$80.0800 \pm 4.4031\%$	$82.2800 \pm 4.8411\%$	$74.7600 \pm 5.5689\%$
Small CK vs. No CK	$97.4000 \pm 2.0000\%$	$95.3200 \pm 2.3306\%$	$97.2000 \pm 1.9280\%$

nuclear size to be considered as a candidate for an HD-CTC but low presentation of CK. For this reason we would not expect to be able to perfectly separate these two classes. A similar argument could be made for Small CK which is a population of cells with HD-CTC level CK expression but small nuclear size. We would, therefore, expect to be able to separate Small Ck from No Ck with high levels of accuracy due to the fact that these marginal populations exist on opposite ends of the HD-CTC spectrum. The term "apoptotic" means dying and refers to a population of cells that are unhealthy or dying. Apoptotic cells are considered to be of interest because they are distinct from WBCs, appear similar in some ways to HD-CTCs, and may suggest some measure of sickness in a patient. Also, Apopotic cells share some similarities with each of the other high interest classes and we would therefore anticipate a greater challenge to separate Apoptotic cells from each other population. Thus, the majority of our results are consistent with the thoughts of trained technicians. Perhaps the most interesting of the pairwise results is the significantly lower ability to separate Apoptotic from Small CK. While the classification rate is far above random, the results of this pairwise classification may result in the formulation of additional biologically answerable questions. For example, in a genomic analysis are there more genes shared between Apoptotic and Small CK cells than other marginal populations? This particular pairwise classification also inspires us to, in future work, test the performance of a non-linear classifier on the same data.

The results in the previous two tables were generated on small sample sizes from the curated data since there were only 100 cells of each marginal population put into the curated data set. However, because we do not need segmentation, only a cell center, one could use the database of EOHI from TKL to run FRD descriptor on a larger uncurated data set.

TABLE 5.6. Results of FRD on Large Uncurated Data Set on Pairwise Classification of EOHI Types

	Mean-Centered and Scaled	Unaltered
HD-CTC vs. Apoptotic	$81.3520 \pm 1.8421\%$	$82.2326 \pm 1.9762\%$
HD-CTC vs. No CK	$86.7720 \pm 1.7846\%$	$87.5320 \pm 1.8309\%$
HD-CTC vs. Small CK	$89.6560 \pm 1.6818\%$	$90.6086 \pm 1.7352\%$
Apoptotic vs. No CK	$89.4840 \pm 1.5542\%$	$89.7200 \pm 1.7207\%$
Apoptotic vs. Small CK	$64.9000 \pm 2.5082\%$	$66.2600 \pm 2.7072\%$
Small CK vs. No CK	$93.7200 \pm 1.2384\%$	$92.9040 \pm 1.3461\%$

The values in Table 5.6 are the average accuracy over 500 trials performed on an uncurated data set consisting of 2000 cells, 500 cells of each high interest class. There was no hand selection involved in the selection of these cells. In each trial 75% of each class is used for training and the remaining 25% is used for testing. Other image representations were not run on this large EOHI data set because of their dependence on segmentation and the inability to generalize the methods to uncurated data. We see in Table 5.6 that when we generalize FRD to uncurated data on the pairwise classification tasks we loose a significant amount of performance in each classification task, especially in the Apoptotic versus Small CK problem. The reduction in performance suggests that exploration into non-linear SVM may be worthwhile to see if the classifiers are more robust on the pairwise problems.

CHAPTER 6

Conclusions and Future Work

The analysis provided in Chapter 5 leads us to conclude that the non mean-centered FRD is the best of the three proposed descriptors for the EOHI versus WBC classification task. This conclusion is based on the ultimate goal of being able to implement our algorithm on a large scale while minimizing the number of HD-CTCs that are misclassified. Although there is no statistically significant difference in the average accuracy of the three methods on the data set considered here, the ability of FRD to be applied to images with only a center of mass computation on the nuclear channel required and its rotational invariance makes it superior for the current application. The results of work to date also suggests that FRD produces effective low level feature descriptors for separating EOHI from one another.

Although currently we are not able to implement our methods on full patient slides in a time which meets the goals of our original problem statement, there are many things we can do to improve the performance. For example, we have not explored the minimum number of points required to be sampled on a ring to determine appropriate classification. We note that these algorithms are trivially parallelized and substantial accelerations in the code are possible. It should be informative to explore the levels of sensitivity and specificity of the descriptor when applied to complete slides.

Given the promising results of this first exploration we are hopeful that the low level features that we can determine for HD-CTC, and near HD-CTC, events will provide further treatment information. Future work will include applying FRD to other circulating epithelial cell (CEC) classification tasks including, but not limited to, differentiation of HD-CTCs in

patients with known malignant tumors from CEC in patients with benign masses, differentiation of pre and post biopsy cells, and differentiation of cells at different times before and after various treatments. Due to the significant difficulty in many of these future explorations we will also explore using a non-linear kernel SVM classifier.

All of our analysis has been done, to date, on images generated using a 10x objective. Future work will include testing the performance of the FRD on images produced using higher magnification. Generating higher magnification images results in too many images to effectively store when several image channels are used. It may be possible to show that a descriptor performs with comparable or improved classification rates on a single channel only at higher magnification, thereby reducing the need for the additional image channels and allowing greater textural variation. There is hope that with higher levels of textural information we may be able to make the automated detection of HD-CTC using FRD a tool in cancer diagnostics.

Another problem of future interest is the unsupervised cell subtype detection. The unsupervised cell classification problem refers to determining robust cell classes that can be strongly differentiated mathematically but do not have known clinical titles. It is our hope that by combining both the supervised and unsupervised cell types that we may be able to connect a patient's relative composition of all cell types to their clinical status at the time the sample was acquired.

Additionally, we would like to see if there is a way to establish a quantifiable difference between cell line circulating tumor cells and HD-CTCs as found in patient blood to further support the findings of The Kuhn Laboratory published in [24]. This is an important concept because much of the cancer research being done currently is done using cell lines and not cells

developed within patients and therefore the results of a great deal of therapeutic research may not be transferable to patients.

Bibliography

[1] Dena Marrinucci, Kelly Bethel, Anand Kolatkar, Madelyn Luttgen, Michael Malchiodi, Franziska Baehring, Katharina Voigt, Daniel Lazar, Jorge Nieva, Lyudmila Bazhenova, Andrew Ko, Michael Korn, Ethan Schram, Michael Coward, Xing Yang, Thomas Metzner, Rachelle Lamy, Meghana Honnatti, Craig Yoshioka, Joshua Kunken, Yelena Petrova, Devin Sok, David Nelson, and Peter Kuhn. Fluid biopsy in patients with metastatic prostate, pancreatic and breast cancers. *Physical Biology*, 9(1):016003, 2012.

[2] Mario Giuliano, Antonio Giordano, Summer Jackson, Kenneth R Hess, Ugo De Giorgi, Michal Mego, Beverly C Handy, Naoto T Ueno, Ricardo H Alvarez, Michelino De Laurentiis, et al. Circulating tumor cells as prognostic and predictive markers in metastatic breast cancer patients receiving first-line systemic treatment. *Breast Cancer Res*, 13(3): R67, 2011.

[3] SJ Cohen, CJA Punt, N Iannotti, BH Saidman, KD Sabbath, NY Gabrail, J Picus, MA Morse, E Mitchell, MC Miller, et al. Prognostic significance of circulating tumor cells in patients with metastatic colorectal cancer. *Annals of oncology*, 20(7):1223–1229, 2009.

[4] Mohid S Khan, Amy Kirkwood, Theodora Tsigani, Jorge Garcia-Hernandez, John A Hartley, Martyn E Caplin, and Tim Meyer. Circulating tumor cells as prognostic markers in neuroendocrine tumors. *Journal of Clinical Oncology*, 31(3):365–372, 2013.

[5] Jorge Nieva, Marco Wendel, Madelyn S Luttgen, Dena Marrinucci, Lyudmila Bazhenova, Anand Kolatkar, Roger Santala, Brock Whittenberger, James Burke, Melissa Torrey, et al. High-definition imaging of circulating tumor cells and associated cellular events in non-small cell lung cancer patients: a longitudinal analysis. *Physical biology*, 9(1):016004, 2012.

[6] Daniel C Danila, Glenn Heller, Gretchen A Gignac, Rita Gonzalez-Espinoza, Aseem Anand, Erika Tanaka, Hans Lilja, Lawrence Schwartz, Steven Larson, Martin Fleisher, et al. Circulating tumor cell number and prognosis in progressive castration-resistant prostate cancer. *Clinical Cancer Research*, 13(23):7053–7058, 2007.

[7] Katharina Pachmann, Oumar Camara, Andreas Kavallaris, Sabine Krauspe, Nele Malarski, Mieczyslaw Gajda, Torsten Kroll, Cornelia Jörke, Ulrike Hammer, Annelore Altendorf-Hofmann, et al. Monitoring the response of circulating epithelial tumor cells to adjuvant chemotherapy in breast cancer allows detection of patients at risk of early relapse. *Journal of Clinical Oncology*, 26(8):1208–1215, 2008.

[8] Sabine Riethdorf, Herbert Fritsche, Volkmar Müller, Thomas Rau, Christian Schindlbeck, Brigitte Rack, Wolfgang Janni, Cornelia Coith, Katrin Beck, Fritz Jänicke, et al. Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system. *Clinical Cancer Research*, 13(3):920–928, 2007.

[9] David Hopkins. A computer vision approach to classification of circulating tumor cells. Master's thesis, Colorado State University, Fort Collins, Colorado, 2012.

[10] Percannella, Fogia, and Soda. 1st international contest on HEp-2 cells classification, November 2012. URL `http://mivia.unisa.it/hep2contest`.

[11] Jan Oosterwijk, Cecile Knepfl, Wilma Mesker, Hans Vrolijk, Willem Sloos, Hans Pattenier, Ilya Ravkin, and Gert-Jan van Ommen. Strategies for rare-event detection: An approach for automated fetal cell detection in maternal blood. *The American Journal of Human Genetics*, 63(6):1783 – 1792, 1998. ISSN 0002-9297. doi: 10.1086/302140. URL `http://www.sciencedirect.com/science/article/pii/S0002929707616243`.

[12] Jaana Karttunen, Sarah Sanderson, and Nilabh Shastri. Detection of rare antigen-presenting cells by the lacZ T-cell activation assay suggests an expression cloning strategy for T-cell antigens. *Proceedings of the National Academy of Sciences*, 89(13):6020–6024, 1992.

[13] Stine-Kathrein Kraeft, Rebecca Sutherland, Laura Gravelin, Guan-Hong Hu, Louis H Ferland, Paul Richardson, Anthony Elias, and Lan Bo Chen. Detection and analysis of cancer cells in blood and bone marrow using a rare event imaging system. *Clinical cancer research*, 6(2):434–442, 2000.

[14] Jean-Yves Pierga, Charlyne Bonneton, Anne Vincent-Salomon, Patricia de Cremoux, Claude Nos, Nathalie Blin, Pierre Pouillart, Jean-Paul Thiery, and Henri Magdelénat. Clinical significance of immunocytochemical detection of tumor cells using digital microscopy in peripheral blood and bone marrow of breast cancer patients. *Clinical Cancer Research*, 10(4):1392–1400, 2004.

[15] Giovanni Pauletti, Suganda Dandekar, HongMei Rong, Lilllian Ramos, HongJun Peng, Ram Seshadri, and Dennis J Slamon. Assessment of methods for tissue-based detection of the HER-2/neu alteration in human breast cancer: a direct comparison of fluorescence in situ hybridization and immunohistochemistry. *Journal of Clinical Oncology*, 18(21): 3651–3664, 2000.

[16] R.M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-3(6):610–621, 1973. ISSN 0018-9472. doi: 10.1109/TSMC.1973.4309314.

[17] Madelyn Luttgen. Personal Communication.

[18] Rico Hiemann, Thomas Büttner, Thorsten Krieger, Dirk Roggenbuck, Ulrich Sack, and Karsten Conrad. Challenges of automated screening and differentiation of non-organ specific autoantibodies on HEp-2 cells. *Autoimmunity Reviews*, 9(1):17, 2009.

[19] Yu-Len Huang, Yu-Lang Jao, Tsu-Yi Hsieh, and Chia-Wei Chung. Adaptive automatic segmentation of HEp-2 cells in indirect immunofluorescence images. In *Sensor Networks, Ubiquitous and Trustworthy Computing, 2008. SUTC'08. IEEE International Conference on*, pages 418–422. IEEE, 2008.

[20] Michael Kirby. *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns*. John Wiley & Sons, Inc., 2000.

[21] Jason D Hipp, Jerome Y Cheng, Mehmet Toner, Ronald G Tompkins, Ulysses J Balis, JD Hipp, JY Cheng, M Toner, RG Tompkins, UJ Balis, et al. Spatially invariant vector quantization: A pattern matching algorithm for multiple classes of image subject matter including pathology. *Journal of pathology informatics*, 2:13, 2011.

[22] Arnold Wiliem, Yongkang Wong, Conrad Sanderson, Peter Hobson, Shaokang Chen, and Brian C. Lovell. Classification of human epithelial type 2 cell indirect immunofluoresence images via codebook based descriptors. In *IEEE Workshop on Applications of Computer Vision*. IEEE, 2013.

[23] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[24] Daniel C Lazar, Edward H Cho, Madelyn S Luttgen, Thomas J Metzner, Maria Loressa Uson, Melissa Torrey, Mitchell E Gross, and Peter Kuhn. Cytometric comparisons between circulating tumor cells from prostate cancer patients and the prostate-tumor-derived LNCaP cell line. *Physical biology*, 9(1):016002, 2012.

APPENDIX A

CODE

1.1. Code Used to Stretch and Shrink Boundaries

```
function [distorted_boundary] = distortion( boundary, center,amount  )


%STRETCH takes in a boundary and will shift each point onvthe boundary by "amount"
    %The following if statements add or subtract the distortion amount
    %depending on the center of the cell. A negative distortion amount
    %corresponds to shrinking of the boundary while a positve corresponds
    %to stretching.


for i=1:length(boundary)
    x=boundary(i,1);
    y=boundary(i,2);
        if x>center(1,1)
            x_stretch=x+amount;
        elseif x<=center(1,1)
            x_stretch=x-amount;
        end
        if y>center(1,2)
            y_stretch=y+amount;
        elseif y<=center(1,2)
            y_stretch=y-amount;
```

```
        end
    distorted_boundary(i,1)=x_stretch;
    distorted_boundary(i,2)=y_stretch;
end
```

1.2. Code Used to Build FRD

```
function [points]=circles(x,y,r,np)
% CIRCLES takes in a center (x,y) and a radius r and determines the
% coordinates of number of points (np) along the circle.
angle_step=(2*pi/(np));
ang=0:angle_step:(2*pi)-angle_step;
xp=r*cos(ang);
yp=r*sin(ang);
xp=x+ xp';
yp=y+ yp';
points=[xp yp];
function [descriptor] = circular_descriptor(x,y,r,np,Image)
%CIRCULAR DESCRIPTOR takes in a center (x,y), a radius r, a number of
%points along the circle np, and an Image and interpolates the image value
%at the points along the circle.
[points]=circles(x,y,r,np);
ZI=uint8(interp2(Image, points(:,1), points(:,2), 'cubic'));
descriptor=ZI;
end
```

```
function [ frd_descriptor ] = FRD( image, center, num_rings )
%CRFFT computers the concentric ring fourier transform feature descriptor
%of an image given a center and the number of rings
x=center(1,1);
y=center(1,2);
for r=1:num_rings
    np=r*8;
    [descriptor] = circular_descriptor(x,y,r,np,image);
    fourier_coefficients{1,r}=abs(fft(descriptor))';
end
crfft_descriptor=cell2mat(fourier_coefficients);
end
```

1.3. Code Used to Generate Results of Random Linear SVM Classifiers

```
function [ results ] = random_svm(descriptor_matrix, num_trials, num_each_class )
%RANDOM_SVM takes in a matrix of descriptors where the first half of the
%matrix comes from one class and the last half the second type.
%   Uses LIBLINEAR matlab imterface to generate num_trials classifiers
%   where the 75% of data used for training is randomly selected by
%   permuting the number of events in each class. The RESULTS matrix
%   contains the overall accuracy, the indicies of events that were
%   misclassified, the events used in training and testing, the model
%   determined by LIBLINEAR and the time taken to produce the classifier.
results=cell(num_trials,6);
```

```
for i=1:num_trials
    tic
    clear ind_class_1 ind_class_2
    ind_class_1=randperm(num_each_class);
    ind_class_2=randperm(num_each_class);
    z=floor(.75*num_each_class)
    results{i,3}=[ind_class_1(1,1:z) num_each_class+ind_class_2(1,1:z)]';
    results{i,4}=[ind_class_1(1,z+1:end) num_each_class+ind_class_2(1,z+1:end)]';
    train_class_1=descriptor_matrix(ind_class_1(1:z),:);
    train_class_2=descriptor_matrix(num_each_class+ind_class_2(1:z),:);
    rand_train_set=[train_class_1; train_class_2];
    rand_train_label=[ones(z,1);-1*ones(z,1)];
    test_class_1=descriptor_matrix(ind_class_1(z+1:end),:);
    test_class_2=descriptor_matrix(num_each_class+ind_class_2(z+1:end),:);
    rand_test_set=[test_class_1; test_class_2];
    rand_test_label=[ones(num_each_class-z,1);-1*ones(num_each_class-z,1)];

    % The train and predict functions are from LIBLINEAR

    model = train(rand_train_label, sparse(double(rand_train_set)),...
     ...['-s 5','-c .5']);
```

```
    [predict_label, accuracy, dec_values] = predict(rand_test_label,...
     ... sparse(double(rand_test_set)), model)
    results{i,1}=accuracy(1,1);
    s=find(predict_label~=rand_test_label);
    indicies=[ind_class_1(1,z+1:end) 500+ind_class_2(1,z+1:end)]';
    misclassified=indicies(s,:);
    results{i,2}=misclassified;
    results{i,5}=model.w;
    toc
    results{i,6}=toc;
end
end
```