Integrated multiparametric deep spatial phenotyping of mouse models of lung adenocarcinoma

by

Yang Dai

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author . . . .
Department of Electrical Engineering and Computer Science
September 3, 2019

Certified by . . . .
Sandro Santagata
Associate Professor in Pathology, Harvard Medical School
Thesis Supervisor

Certified by . . . .
Tyler Jacks
Professor of Biology, MIT
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Department Committee on Graduate Theses

Integrated multiparametric deep spatial phenotyping of mouse models of lung adenocarcinoma

by

Yang Dai

Submitted to the Department of Electrical Engineering and Computer Science
on September 3, 2019, in partial fulfillment of the
requirements for the degree of
Masters of Engineering in Computer Science and Engineering

Abstract

In this thesis, I developed computational pipelines and algorithms that use high dimensional biomarker imaging data to predict features of tumor tissues taken from a genetically engineered mouse model (GEMM) of lung adenocarcinoma. I extracted biomarker expression levels and morphological, textural, and spatial motifs of single cells from the imaging data and used these features to train algorithms to predict tumor histologic grade, a measure correlated with the malignant potential of a tumor. The algorithm predictions were evaluated through comparison to a validated deep learning model. The random forest algorithm achieved a 72% accuracy classifying cells as belonging to a non-tumor, grade 1, grade 2, or grade 3 region and achieved a 87% accuracy classifying cells as belonging to a tumor or non-tumor region. A combination of biomarker, morphological, textural, and spatial features generated models that performed better than any single group of markers by itself; spatial features in particular significantly improved model performance.

Thesis Supervisor: Sandro Santagata
Title: Associate Professor in Pathology, Harvard Medical School

Thesis Supervisor: Tyler Jacks
Title: Professor of Biology, MIT

Acknowledgments

There are many people I want to thank without whom this thesis would not have been possible. First I would like to thank my co-advisor, Dr. Santagata, who guided me throughout the process, read drafts, and provided valuable feedback. He has allowed me to work on interesting projects and has guided me in learning both biology and computational methods. I would like to thank my other co-advisor, Dr. Jacks, for his support on the thesis and his lab for their continued collaboration. They not only shared their tissue samples but also provided the tumor grading used in this project. I would also like to thank Giorgio, my mentor in the lab, who guided me step-by-step through the data collection and processing pipeline and provided valuable feedback to improve the methods and algorithms in this thesis.

Dr. Sorger and members and staff at the Lab for Systems Pharmacology provided valuable support and generously shared both their lab space and instruments. I would also like to thank Walid, a member of the Agar lab, for his guidance on image registration.

This project could not have been done without the day-to-day support from all of my labmates, Claire, Ru, Ziming, and Danae, who have basically taught me everything I know about working in a wet lab and made the lab an enjoyable place to be. Lastly, I would like to thank my roommate and fellow MEng student, Gina, who has kept me on top of deadlines and provided entertaining company during our thesis writing sessions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the recent development of multiplexed technologies for tissue imaging, investigators can generate multidimensional and spatially resolved biological data at single cell resolution [12]. Managing and processing this wealth of information requires efficient data workflows and pipelines. A key analytic step is the automated detection of cell subpopulations and the measurement of biomarker expression levels within cells. Such information not only facilitates hypothesis generation, but it also permits the testing of associations between biomarker expression and other morphologic and functional features of interest.

The high dimensionality of the data (e.g. biomarker expression levels, patterns of expression, intracellular localization, co-expression of biomarkers and morphologic features) has generated great potential for leveraging the power of machine learning algorithms to extract patterns that can facilitate biomedical research. In particular, it has allowed researchers to make observations about the development and progression of cancer [12].

Lung adenocarcinoma is the most prevalent form of lung cancer in the U.S. and is the leading cause of cancer deaths worldwide [9]. To better understand the development and mechanisms of the disease, Tyler Jacks' lab at MIT has created a genetically engineered mouse model (GEMM) of lung cancer that allows for controlled timing and multiplicity of tumor development and also allows for recapitulation of the genetic altertions found in the human version of the disease. This model allows researchers to monitor the progression of tumors and to draw comparisons with mechanisms of lung tumorigenesis in humans. In my thesis work, I used lung tissues from this GEMM model.

I studied lung tissues from this disease model using Haemotoxylin and Eosin (H&E) staining and tissue-based cyclic immunofluorescence (t-CyCIF). H&E stained tissue sections are used to identify tissue types and morphological changes which are integral to the diagnosis of cancer [5]. For this project, a company specializing in deep learning models for imaging data processed the H&E stained slides to detect tumor regions and their corresponding histologic grades. t-CyCIF is a method for fluorescence imaging developed in the Lab for Systems Pharmacology (LSP) that can be used to detect the expression levels of up to 60 biomarkers on a single slice of tissue while preserving the spatial arrangement of the cells within the tissue. These methods provide information about the expression levels and locations of biomarkers as well as histologic features associated with tumors such as their grade.

Chapter two describes the t-CyCIF imaging procedure, the pipeline to obtain single-cell data from the images, and the pipeline to integrate tumor grading data with t-CyCIF single cell data.

Chapter three describes the single cell analysis methods used to classify individual cells and methods to extract morphological and textural features of cells.

Chapter four describes the algorithms used to predict tumor grade and the evaluation of these algorithms.

Chapter five discusses the performance of the algorithms, the biological interpretation of the results, and potential extensions of this project.

1.1 Background

1.1.1 Mouse model of lung cancer

The mouse model in this project uses conditional activation of the K-ras oncogene and loss of function of p53 in the lungs of mice (KP mouse model) [9]. K-ras is an oncogenic protein that regulates cell proliferation, differentiation, and survival. Mutations that inactivate K-ras drive the development of lung adenocarcinoma in humans. A quarter to one half of human lung adenocarcinomas and greater than 90% of mouse lung adenocarcinomas (both spontaneous and chemically induced) are found to have activating mutations in K-ras. p53 is a tumor suppressor that induces growth arrest or apoptosis and that negatively regulates cell division.

In the mouse model, LoxP DNA elements surround a "stop" element in front of the oncogenic mutant K-ras G12D and flanks the second and tenth exons of the p53 tumor suppressor gene. Adenoviruses expressing Cre administered intranasally to the mice delivers Cre recombinase to the lung cells, which eliminates the "stop" element in front of the K-ras oncogene and deletes exons two through ten of p53. This induces deletion of K-ras and loss of function of p53 [8].

Additionally, the Jacks lab has developed a CRISPR/Cas9-based approach to investigate gene mutations related to tumorigenesis in the KP mouse model [20]. In this study, I used KP mice chimeric for inactivation of heat shock factor 1 (HSF1), a transcriptional regulator of chaperone gene expression that is thought to play a significant role in cancer progression [24].

1.1.2 t-CyCIF

t-CyCIF is a method that builds upon techniques in immunofluorescence imaging. Direct immunofluorescence imaging utilizes fluorophore conjugated antibodies that are developed to bind to specific epitopes within proteins. Conventional fluorescence microscopes are used to detect the light emitted from the fluorophore conjugated antibodies thereby providing the location and abundance level of protein expression within cells and tissues [12]. This information is collected on adjacent slices of tissue of 5 to 10 micron thickness. The drawback to traditional immunofluorescence imaging is that it allows for only one round of imaging per tissue slice and thus restricts the number of different proteins that can be observed per tissue. t-CyCIF overcomes this limitation by using a protocol that allows new antibodies to be applied to the the same tissue slice and re-imaged for up to 15-20 cycles. This method gives researchers a high-dimensional representation of a single tissue slice in which spatial configuration is preserved.

As an active member of the LSP, Sandro Santagata (primary co-mentor) and members of his lab have implemented this optical imaging method which allows measurements at single cell resolution and readily permits the detection of events occurring at the sub-cellular level (e.g. stress foci in the cell nucleus). The technique has been optimized to work with formalin-fixed, paraffin-embedded (FFPE) specimens, which are widely used in pathology departments for cancer diagnosis and for the analysis of tissue phenotypes in mouse models of disease. Such specimens are archived and stored for long periods of time thus permitting retrospective analysis of precious and sometimes rare human and mouse tissues.

1.2 Related work

1.2.1 Machine learning approaches to single cell classification

Machine learning techniques have been applied to multi-dimensional single cell data to facilitate the detection of cellular subpopulations [5]. These subpopulations can be used to better define condition-specific behaviors of cells that can serve as markers of disease status and predict clinical outcome. Traditional approaches to identifying cellular subpopulations include manual gating, which relies on domain knowledge-driven quantification and thus is labor intensive and difficult to scale to increasingly larger datasets. Other computational methods include nonparametric clustering and density-based methods, but these methods have difficulty estimating the true number of clusters [7].

Bruggner et al. have developed an algorithm, Citrus (cluster identification, characterization, and regression), to automatically identify and stratify subpopulations of cells in multidimensional mass cytometry data [7]. Mass cytometry is a technique that is similar to flow cytometry but rather than using antibodies conjugated to fluorophores to characterize cells that are dissociated into single cell suspensions, mass cytometry antibodies are labeled with heavy metal ion tags and measured using time-of-flight mass spectrometry. With mass cytometry, greater than 40 concurrent parameter measurements can be achieved at a single cell level, but unlike t-CyCIF, the method does not preserve the spatial information about the cells. The data input for Citrus is samples of cells and their corresponding measurements. Each sample is annotated with metadata about the specific patient from which the sample was acquired, the progression of the patient's disease course and the patient's ultimate outcome. The algorithm randomly selects a fraction of cells from all the samples and performs hierarchical clustering on the cells based on marker similarity. Clusters of sufficient size are then used to calculate the cellular features that describe that particular cluster. Features can include the proportion of a sample's cells in each cluster and the median measurements of each functional marker. Lastly, the algorithm uses regularized supervised machine learning to identify features and clusters that best predict a known endpoint, such as clinical outcome. The accuracy of the model is assessed via cross validation using similar sets of samples.

Arvaniti and Claassen developed an algorithm, CellCnn, which combines multiple instance machine learning with convolutional neural networks to identify T cell subsets associated with an increased risk of AIDS onset in a HIV-infected patient cohort. CellCnn was also used to detect rare cell populations associated with minimal residual disease (MRD) in acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The data sets were obtained via either mass cytometry or flow cytometry. CellCnn implements a convolutional neural network which takes as input to its first layer a set of cells and their corresponding measurements. Each measurement is evaluated with respect to each convolutional filter in the convolutional layer, and the pooling layer then takes either the maximum or mean of the results of each convolutional filter. The pooling layer is then connected to the output layer, which contains the classification of the cell. The weights of the layers were optimized using mini-batch stochastic gradient descent with Nesterov momentum [4].

1.2.2 Cell morphology and texture linked to functional states

A review paper by Prasad et al. describes several studies linking cell morphology with functional changes in the cell [17]. Uhler et al. found that changes in cell morphology led to changes in the position of chromosome territories and changes in gene expression. Abnormal cell morphology is already used to aid cancer diagnoses, and quantifiable morphological features could further aid the process. Prasad et al. measured shape features of 8 osteosarcoma cell lines, 4 of which are highly metastatic and 4 of which have low metastatic ability. Using Zernike moments, a rotation-invariant measure of shape, they found 2 types of metastatic cell lines that showed predictive shape changes. Another study found that a neural network could predict metastatic capacity of cell line using morphological markers with 99% accuracy.

In addition to cell morphology, cell texture has also been shown to be predictive of cell properties. Boland et al. developed a method to characterize protein localization patterns using Zernike moments and Haralick features, which measures cell texture. They achieved a 88% accuracy using a backpropagation neural network [6]. Pantic et al. used Haralick features, specifically entropy, angular second moment, correlation, and variance to differentiate thymus cortical lymphocytes and medullar lymphocytes [15]. They found that medullar lymphocytes may have a higher nuclear textural entropy and variance and lower angular second moment and texture correlation than lymphocytes in the thymus cortex.

Chapter 2

Methods

2.1 Tissue collection and imaging

In this project, I used tissues from the lungs of a KP mouse chimeric for inactivation of HSF1. The five lobes of the lung were preserved as FFPE specimens and were imaged using t-CyCIF for 8 cycles, resulting in a total of 18 unique biomarkers. t-CyCIF allows for imaging of up to 3 biomarkers in addition to a blue-fluorescent DNA stain (DAPI) per cycle. In each cycle, separate biomarkers reside in separate frequency channels, which allows the signals to be detected without interference.

The Jacks Lab performed the tissue extraction and preservation, and a member of the Santagata Lab performed the imaging. In addition, H&E staining was applied to adjacent sections of each tissue, and the resulting images were sent to Aiforia to obtain tumor grade data.

2.2 Image processing pipeline

Due to the high resolution of the t-CyCIF images, the image of the whole tissue is typically saved section by section, and these sections are stitched together using an existing program (ASHLAR) to form an image of the whole tissue. ASHLAR also aligns the images from different cycles to account for slight shifts in the tissue that occur over the cycles. Artifacts in the images are corrected using the BaSiC tool [16].

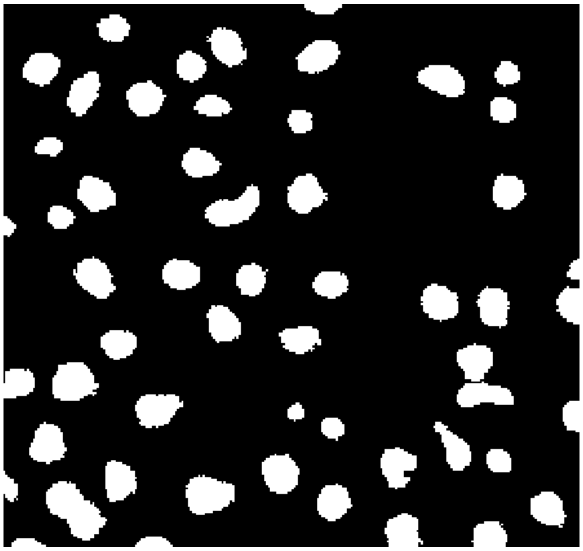


Figure 2-1: Example of segmentation mask.

The next step is to automatically detect individual cells in the image. This is done using an interactive segmentation program, ilastik [22], which outputs a probability mask indicating the probability that a given pixel belongs to the nucleus, cytoplasm, or background. A Matlab script processes the probability mask to output a binary mask separating the image into nuclear areas and background. An example of the binary mask is shown in figure 2-1; the separate nuclear areas (in white) define individual cells. The cytoplasm is defined as the area within 5 pixels of the perimeter of the nuclear area. This segmentation mask is applied to each of the image channels containing the different biomarkers obtains the mean pixel intensity in the nucleus and cytoplasm at a single cell level. We assume that pixel intensity directly correlates with biomarker expression level. Lastly, the segmentation mask is used to extract morphological information about the cells such as area, perimeter, and eccentricity. The output of the image processing pipeline is a matrix in which the rows represent single cells, and the columns represent attributes of the cell such as biomarker expression levels, area, perimeter, and the location of the cell within the tissue.

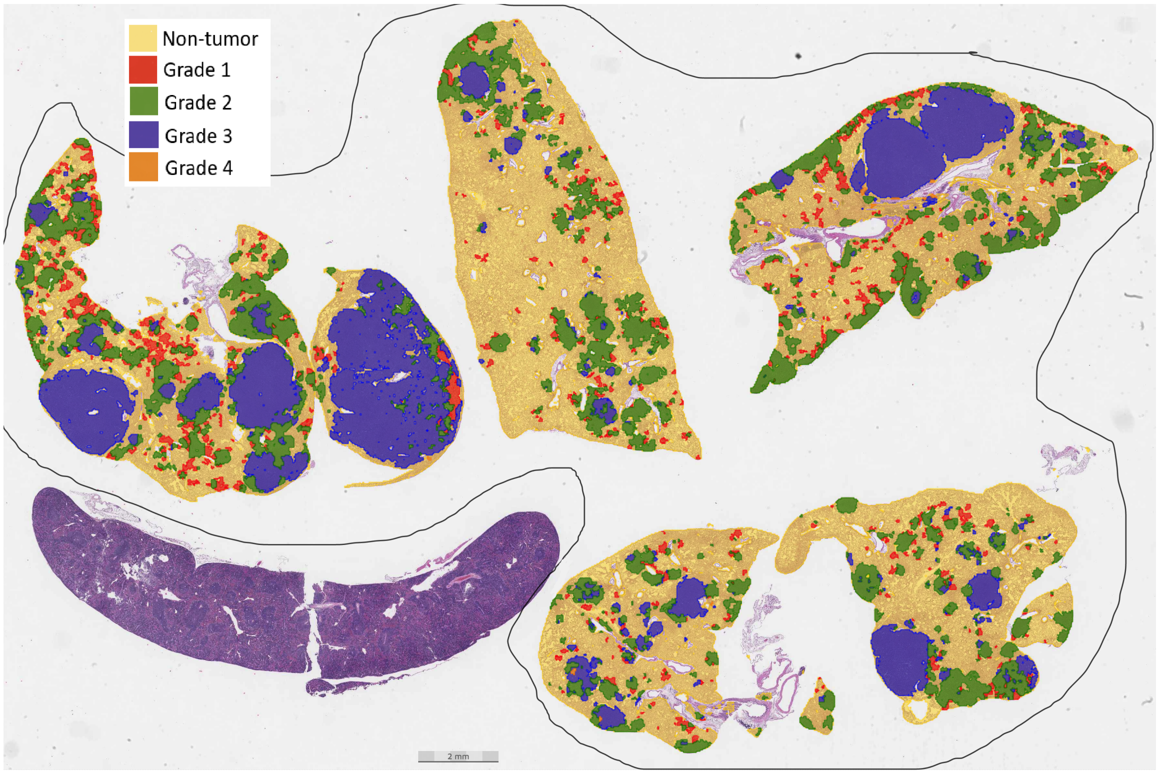


Figure 2-2: Example of Aiforia output with the graded tumor regions overlaid on the H&E stained section.

2.3 Tumor grading

An H&E stained section of the tissue is used to determine the tumor regions and their grades. Aiforia Technologies, a company that builds deep learning models for medical image analysis, processed the H&E images using a trained convolutional neural network (CNN) that categorized regions in the tissue as grade 1, 2, 3, or 4 [1]. Early lesions resembling adenomas are designated as grade 1. Grade 2 tumors are larger adenomas that have slightly enlarged nuclei with prominent nucleoli. Grade 3 tumors are invasive adenocarcinomas with prominent cellular pleomorphism and nuclear atypia and grade 4 tumors are invasive adenocarcinomas with high mitotic index and a distinctive stromal reaction (desmoplasia) [8]. An example of an Aiforia output image is shown in figure 2-2. The yellow regions are normal tissue, the red regions are grade 1 tumors, the green regions are grade 2 tumors, the purple regions are grade 3 tumors, and the dark orange regions are grade 4 tumors.

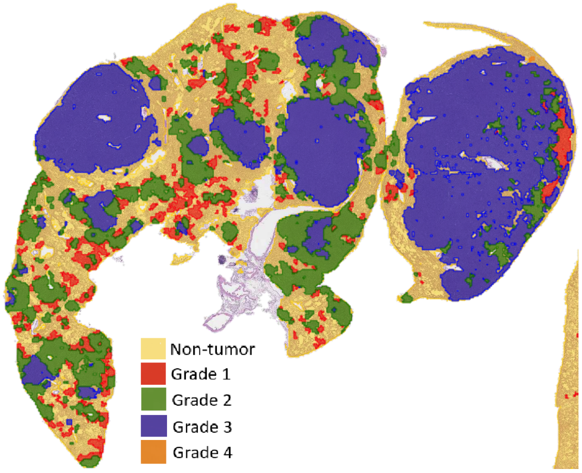


Figure 2-3: Section of one lobe.

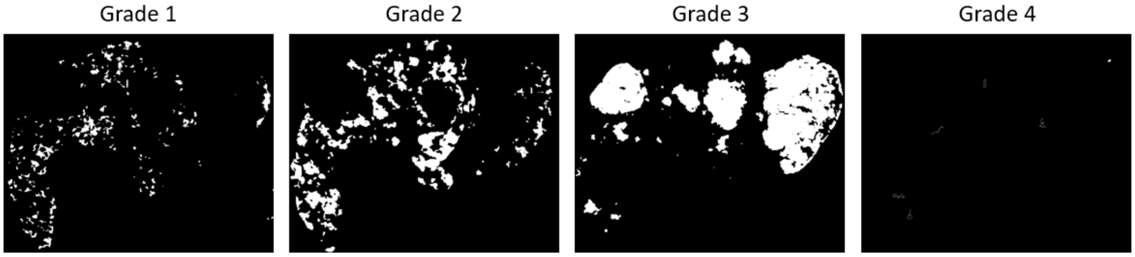


Figure 2-4: Binary masks of tumors grades 1, 2, 3, and 4.

Compared with expert human annotations of an independent dataset, Aiforia's model performed with F1 scores of 89%, 97%, 99%, and 98% for grades 1, 2, 3, and 4, respectively [1]. Given the time-consuming nature of expert human annotations and the high accuracy of the Aiforia model, I used the Aiforia tumor grading as ground truth in this project. Using the difference in colors between the tumor grades, I extracted a binary mask for each tumor grade. There are very few grade 4 tumors because of the relatively early time point at which the lung tissues were harvested, as can be seen in the fourth mask in figure 2-4,

2.4 Image registration

2.4.1 Object detection

Aiforia outputs a lower resolution image and uses a tissue section that is adjacent to the one used in t-CyCIF. It also performs the tumor grading on all five lobes in the same image as opposed to t-CyCIF which images the five lobes separately. This introduces a need for a pipeline to match single cells obtained from t-CyCIF to their corresponding tumor grades obtained from Aiforia.

The object detection pipeline is shown in figure 2-5. The first image is the original output from Aiforia. It was flipped along the x-axis to match the orientation of the tissue in t-CyCIF and converted to grayscale to obtain the second image. The edges were detected using sobel filters (third image) and then dilated to make the boundaries more apparent (fourth image). Finally, the space between the edges were filled to make a binary mask showing the location of the objects (fifth image). A Matlab object detection function was used to detect the lobes from this mask (sixth image). Note that the spleen was not included because we are only interested in the lung tissue for this project.

2.4.2 elastix

The detected objects were matched to their corresponding lobes in t-CyCIF through a process called image registration, which transforms one image to match the other. This process is shown in figure 2-6. I registered the Aiforia image output to the DAPI channel image from t-CyCIF using a toolbox, elastix [11], which is an image registration toolbox specifically designed for medical images. The DAPI image is the fixed image, and the Aiforia image is the moving image, meaning that the Aiforia image is transformed to match the DAPI image. This is because the Aiforia image has much lower resolution. I specified the parameters of the image registration to use only linear transformations to prevent any unnecessary image distortions. An example of the image registration for a single slice of tissue is shown in figure 2-7.

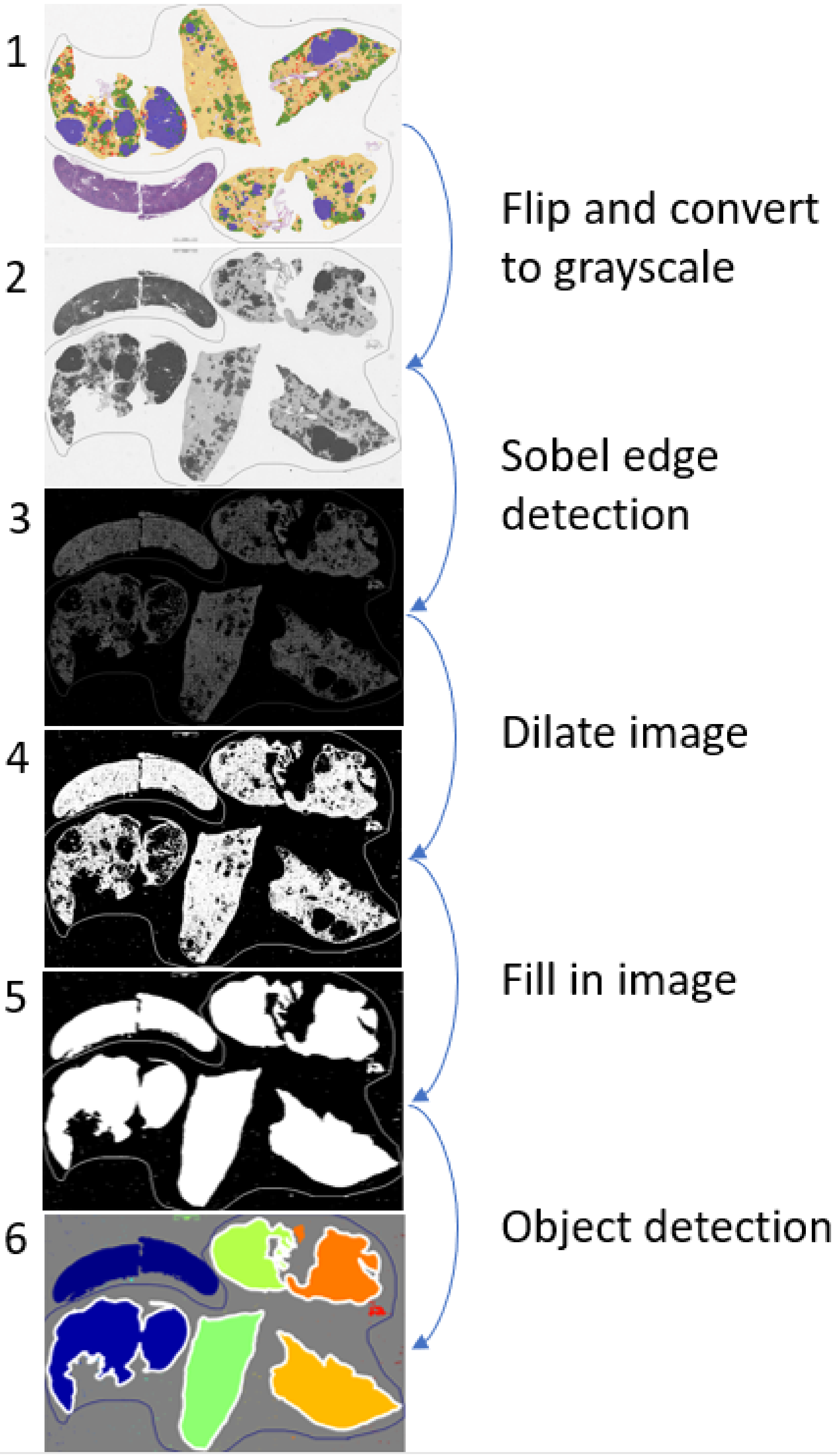


Figure 2-5: Object detection pipeline.

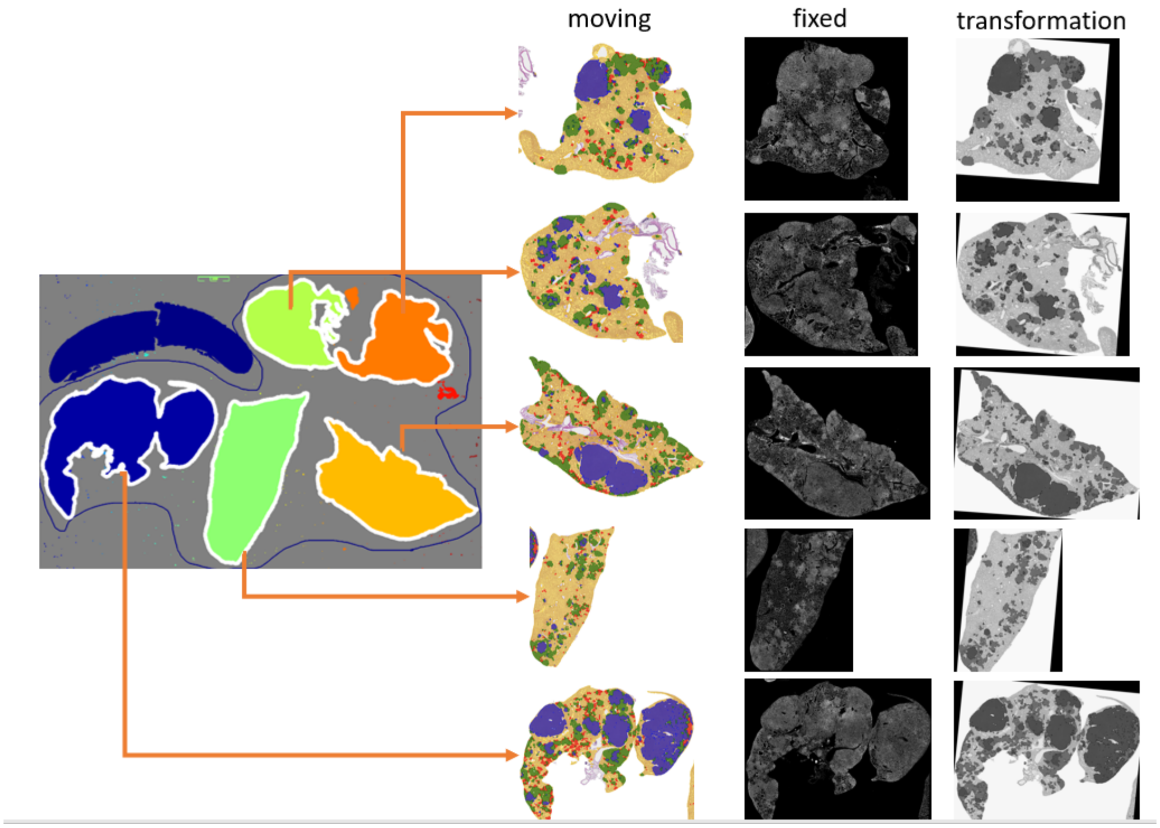


Figure 2-6: Image registration pipeline.

The red is the DAPI image, the green is the transformed Aiforia image, and the third image shows the two images overlaid.

For each image registration, elastix outputs a transformation parameter file. The same transformation file was applied to each of the tumor grade masks extracted from the original image so that the transformed tumor grade masks have the same dimensions as the t-CyCIF image. Figure 2-8 shows the transformed masks (grades 1 to 4) overlaid on the DAPI image. From here, I matched single cells to their corresponding tumor grade using the location of the cell centroid. The tumor grade information was included as a separate column in the single cell data matrix.

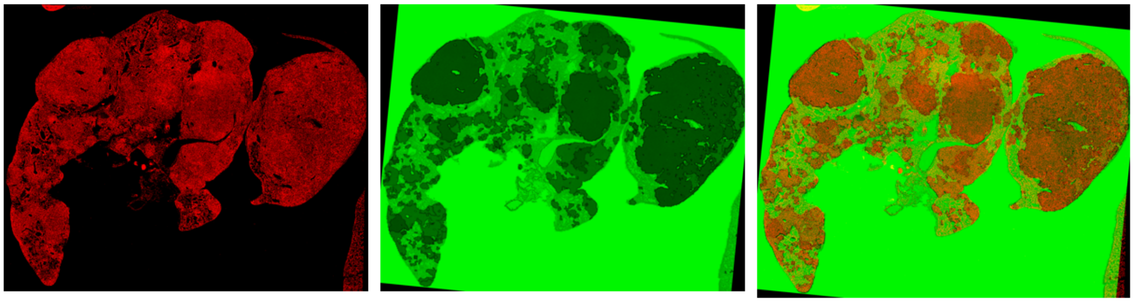


Figure 2-7: Image registration of t-CyCIF and Aiforia grading.

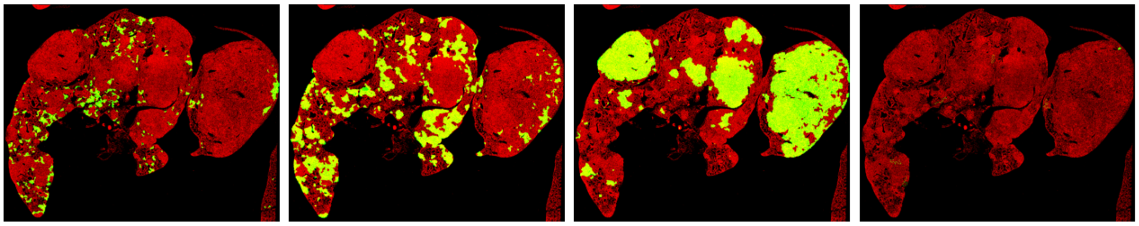


Figure 2-8: Image registration of t-CyCIF and tumor grade masks.

Chapter 3

Single cell analysis

The single cell analysis of the data consists of gating the biomarkers to identify distinct cell populations and extracting morphological, spatial, and textural features of the cells. I used tissue slices from the 5 lung lobes of one mouse. The 5 tissue slices combined and filtered contain a total of 1,622,800 cells. Of these cells, 721,533 (44.5%) belong to non-tumor regions, 81,312 (5.01%) belong to grade 1 tumor regions, 425,144 (26.2%) belong to grade 2 tumor regions, and 394,811 (24.3%) belong to grade 3 and 4 tumor regions regions combined. Grade 4 tumor regions were combined with grade 3 regions because there are very few grade 4 regions.

3.1 Gating of biomarkers

Of the 18 unique biomarkers, 9 showed clear signaling patterns. Descriptions of the biological targets of these markers are shown in table 3.1, and the distributions of the log2 intensities of the biomarker expression levels are shown in figure 3-1. Crops of the fluorescent images for Nkx2.1, Ki67, PCNA, HSF1, CD8, and CD4 overlaid on DAPI (in blue) are shown in figure 3-2.

Certain markers show a bimodal or multi-modal distribution and can be used to differentiate cell sub-populations. In this dataset, there is a clear bimodal distribution for Nkx2.1 and HSF1, and a slight multi-modal distribution for Ki67. CD4 and CD8 also show a slight multi-modal distribution, but the population of CD4 and CD8

Table 3.1: t-CyCIF biomarkers

Biomarker	**Description**
DAPI	Marks nuclear DNA
HSP70	Marks chaperone protein that assists in protein folding [13]
HSP90	Marks chaperone protein that helps regulate proteostatis under stress [21]
HSF1	Marks a protein that is a transcriptional regulator of chaperone gene expression [24]
CD4	Marks helper T cells, which play an important role in the immune system
CD8	Marks cytotoxic T cells, which play an important role in the immune system
PCNA	Marks cell proliferation
Ki67	Marks cell proliferation
Nkx2.1	Marks lung cells

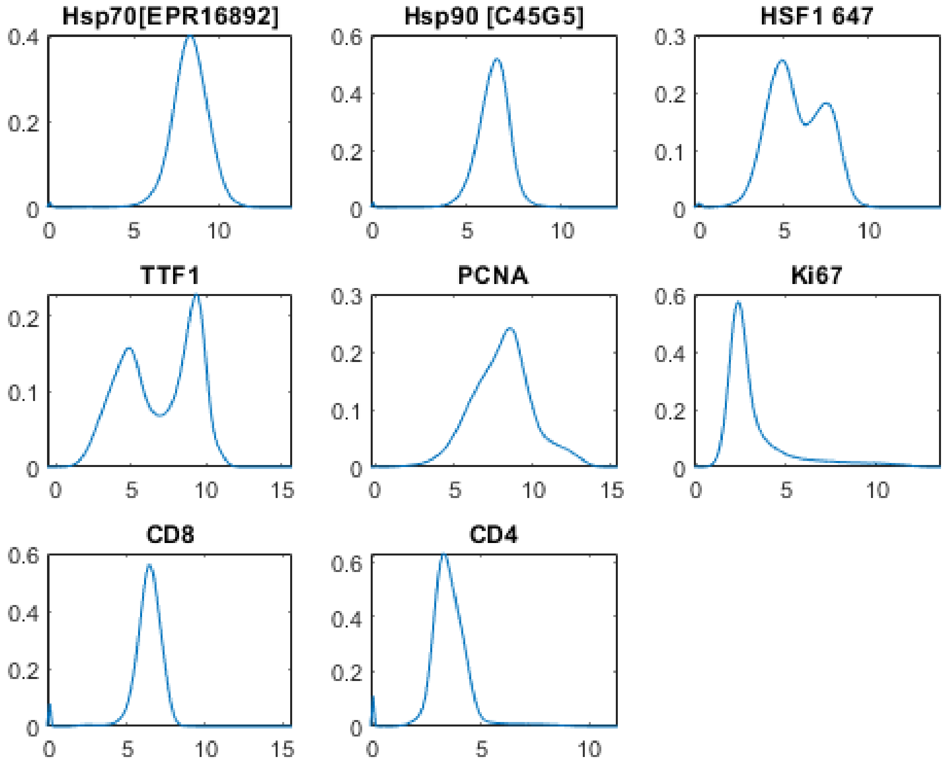


Figure 3-1: Distribution of log2 biomarker intensities

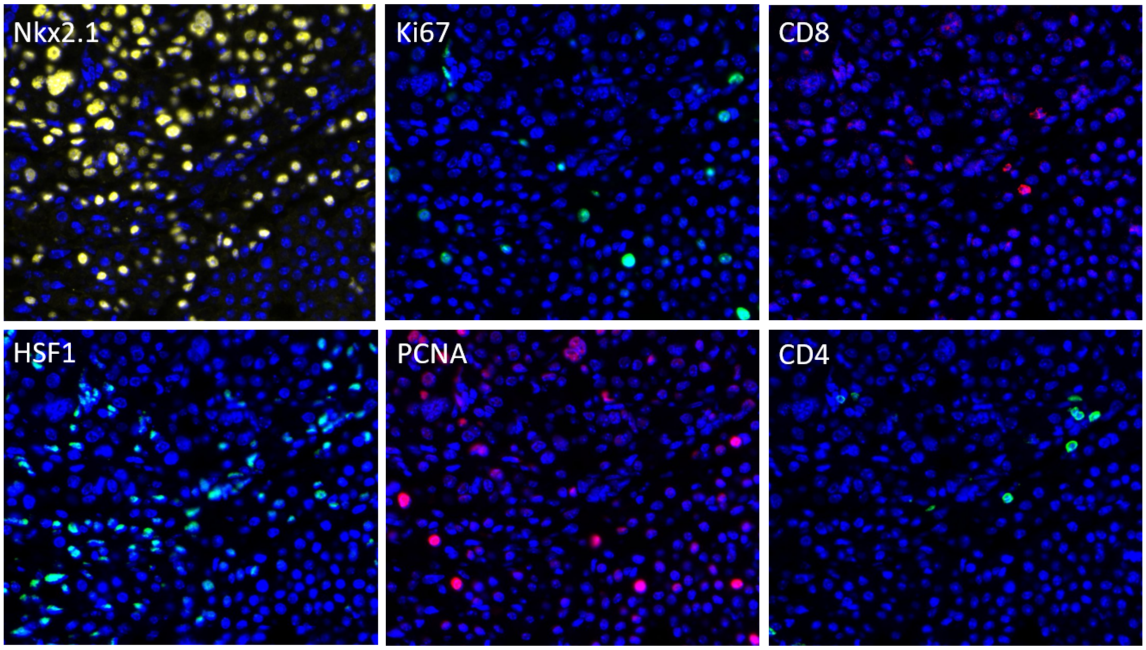


Figure 3-2: Fluorescence images from t-CyCIF.

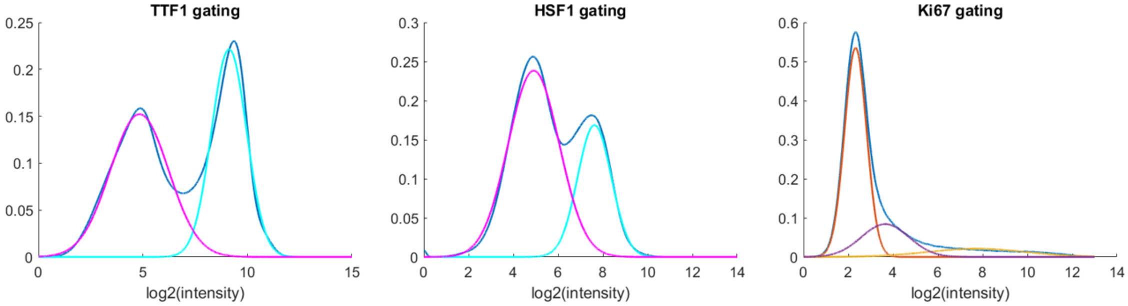


Figure 3-3: Fitting Gaussian mixture models to the biomarker distributions

positive cells is too small to reliably define on the distribution. To define cells that positively express Nkx2.1 or HSF1, I fit a Gaussian mixture model with a cluster size of 2 to the log2 intensity distributions of these markers. I fit a Gaussian mixture model with a cluster size of 3 for Ki67. This is assuming that cells that do not express a certain marker cluster in a normal distribution at a lower intensity value, while cells that positively express that marker cluster in a normal distribution at a higher intensity value. The cutoff is set as the intersection between two normal distributions. The distributions and the gatings are shown in figure 3-3.

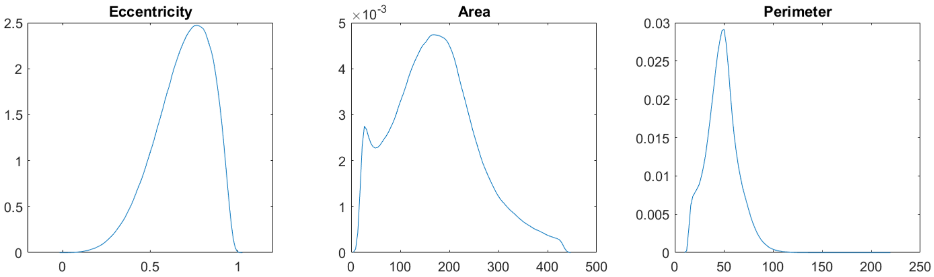


Figure 3-4: Distribution of morphological features.

3.2 Morphological features

The area, perimeter, and eccentricity of the cells are calculated from the segmentation mask, and their distributions are shown in figure 3-4. Eccentricity of an eclipse, a shape that approximates a cell, is defined as

$$\sqrt{(1-\frac{b^2}{a^2})}$$

where a is the length of its semi-major axis, and b is the length of its semi-minor axis. This measure ranges from 0 for a circle and close to 1 for a very elongated eclipse.

Zernike moments are used as quantitative descriptors of cell shape [2] and are calculated using the orthogonal Zernike polynomial basis set. The orthogonality means that there is no redundancy between different moments. The Zernike moment of order n and repetition m for a NxN image is calculated using the following equation:

$$Z_{n,m} = \frac{n+1}{\lambda_n} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) R_{n,m}(\rho_{xy}) e^{-jm}$$

λ_n and ρ_{xy} are normalization factors, and $R_{n,m}$ is a radial polynomial.

In this project, I used the amplitude of the first 3 Zernike moments and their non-negative repetitions; their distributions are shown in figure 3-5. The Zernike moments were calculated using a Matlab function developed by Tahmasbi et. al. [23, 19].

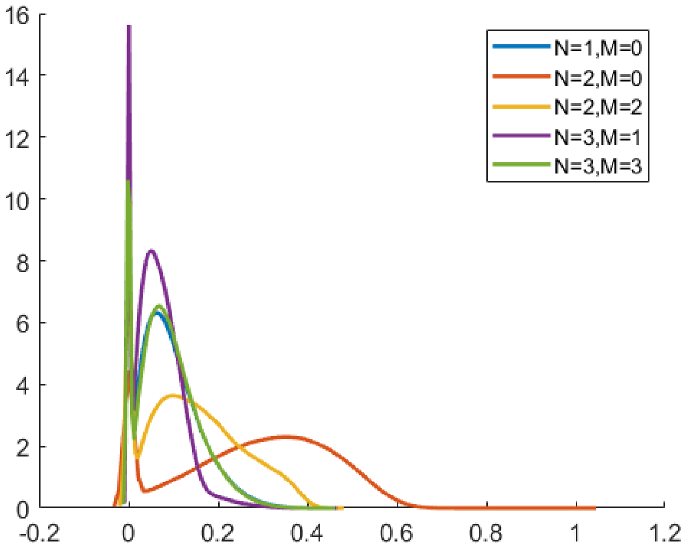


Figure 3-5: Distribution of Zernike moment amplitudes with moment N and repetition M.

3.3 Spatial features

Cell density is directly related to distance between cells; cells in high density regions typically have shorter distances to neighboring cells than cells in sparser areas. I calculated the distances between each cell and its nearest, 10th nearest, and 20th nearest neighbors. Looking at the distance to cells slightly farther away provides a more accurate measure of density. I also calculated the distances to the nearest, 10th nearest, and 20th nearest Nkx2.1 positive cell. Nkx2.1 positive cells are lung cells, so this distance is a measure of lung cell density. Density can be indicative of tumor severity as higher grade tumors tend to consist of more densely packed lung cells [9]. The distance distributions for cell neighbors and lung cell neighbors are shown in figures 3-6 and 3-7 respectively.

3.4 Texture features

Texture is defined as the spatial relationships among gray level values of neighboring pixels. Textural features are extracted from gray level co-occurrence matrices

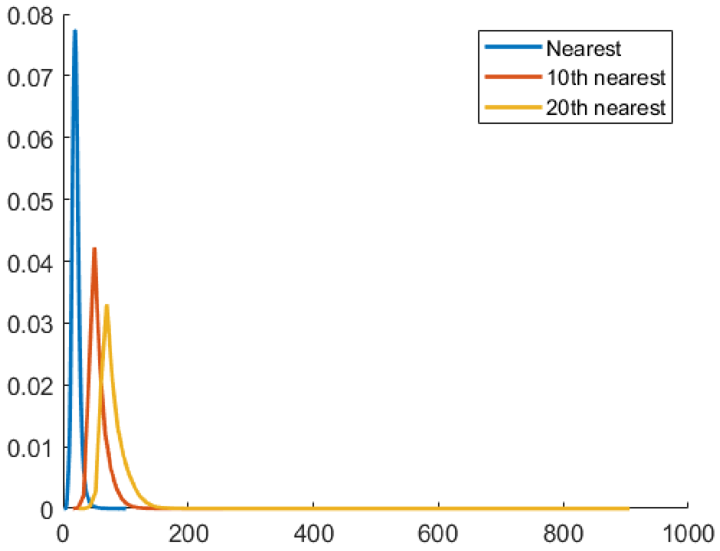


Figure 3-6: Distribution of distances to nearest cells.

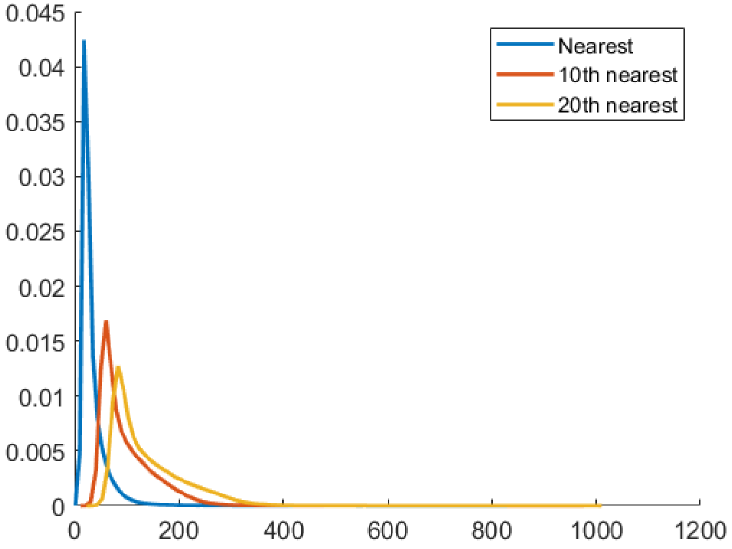


Figure 3-7: Distribution of distances to nearest lung cells.

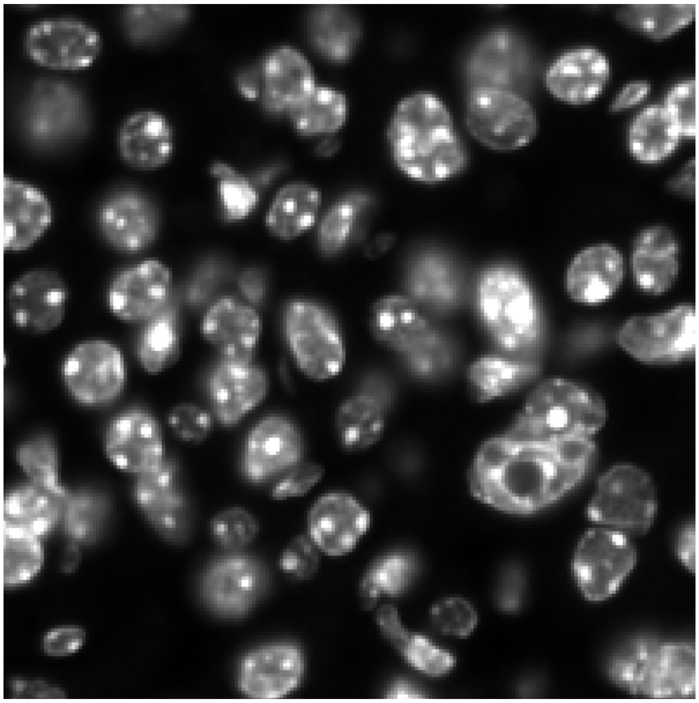


Figure 3-8: Crop of DAPI image in grayscale.

(GLCMs), which are square matrices where the number of rows and columns is equal to the number of gray levels of the image [14]. It represents texture by calculating how often a pair of pixels with specific gray level value in a specified spatial relationship occur in the image. There is one GLCM per spatial relationship. In this project, I used four offsets, one for each neighboring pixel to the top, bottom, left, and right, and took the average of the four values to obtain one GLCM per cell. The cropped images of the cells were taken from the DAPI image converted to grayscale, a crop of which is shown in figure 3-8.

Haralick features extracted from GLCMs are used to quantify textural properties. Of all the Haralick features, entropy, angular second moment (ASM), variance, and correlations are most commonly used in experimental medicine [15]. Given GLCM, P, entropy, ASM, variance, and correlation are defined by the following formulas in which i and j are indices in the matrix, μ is the mean of P and σ is the standard deviation of P.

Entropy:

$$-\sum_i \sum_j P(i,j)log(P(i,j))$$

Angular second moment:

$$\sum_i \sum_j (P(i,j))^2$$

Variance:

$$\sum_i \sum_j (i-\mu)^2 P(i,j)$$

Correlation:

$$\frac{\sum_i \sum_j (ij)P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_x}$$

Entropy is the amount of information needed for image compression. Angular second moment is a measure of homogeneity; it is higher when the pixels are an image are more similar. Variance is a measure of how much pixel values differ from the average, and correlation measures the linear dependency of gray levels of neighboring pixels [14].

The distributions of these textural measures and examples of cells with varying values are shown in figures 3-9, 3-10, 3-11, and 3-12.

3.5 Data cleanup

A significant source of noise in the data is from cell segmentation. The segmentation is not always able to correctly segment cells in high density areas with overlapping cells. There are problems both with over- and under-segmentation; in some cases, background noise is labeled as a cell, while in other cases, multiple cells are segmented as a single cell. Although it is difficult to filter out all incorrect segmentations, there are measures to help determine if a cell is likely the result of a bad segmentation.

To remove cells that are too small or too large and thus likely to be incorrectly segmented, I filtered out cells with an area more than 3 standard deviations away from the median area. Cells with ASM, variance, or correlation equal to 0 and cells with an ASM of 1 were filtered out because these cells are simply a black background.

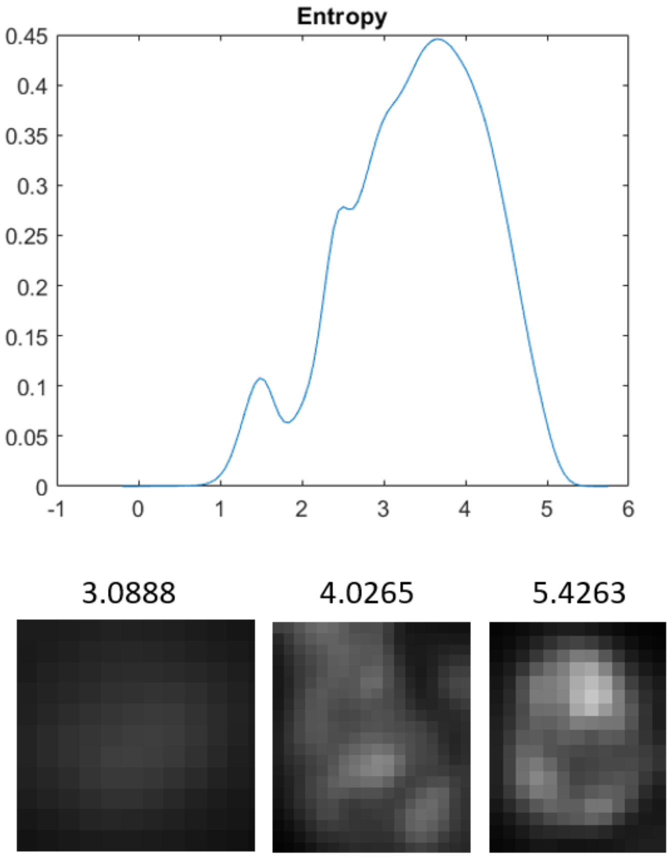


Figure 3-9: Distribution of entropy and examples of different entropy values.

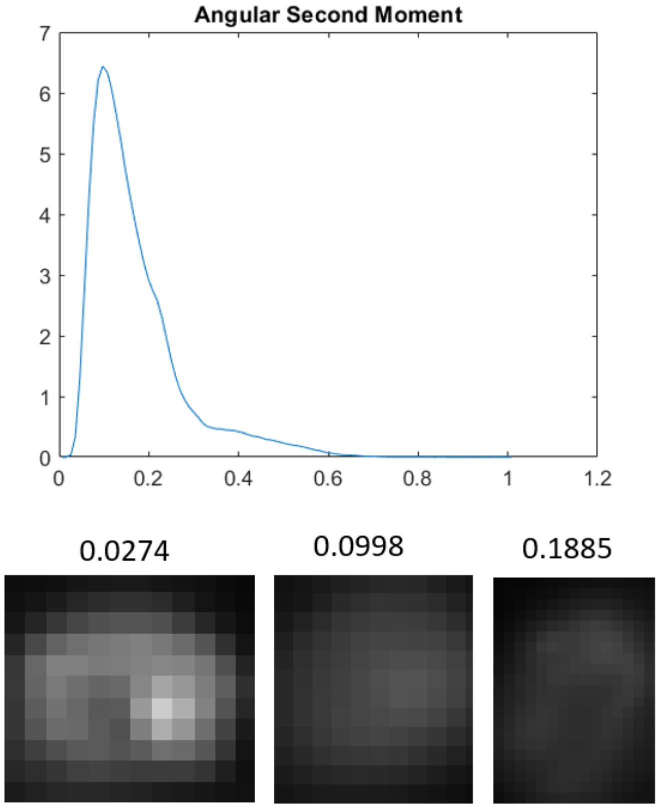


Figure 3-10: Distribution of ASM and examples of different ASM values.

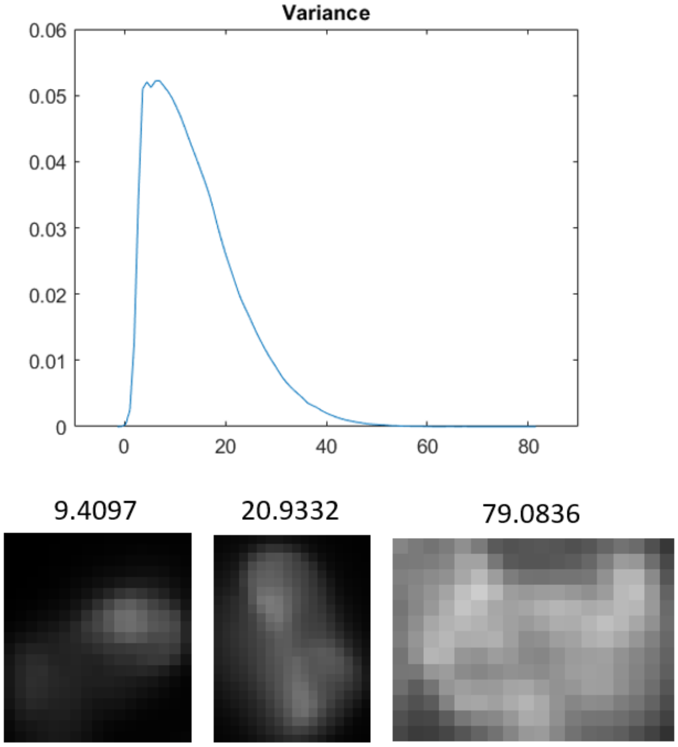


Figure 3-11: Distribution of variance and examples of different variance values.

Out of the 1.6 million cells, a total of 59,014, or 3.41%, were filtered out. Table 3.2 shows the specific filters used and the percentage of cells it filters out.

Table 3.2: Filters for data cleanup

Filter	**% of cells filtered**
Outlier areas	2.30%
ASM, Correlation, Variance, or Entropy = 0	0.82%
ASM = 1	0.39%
Any NaN value	0.54%
Nearest neighbor dist > 200	0.20%

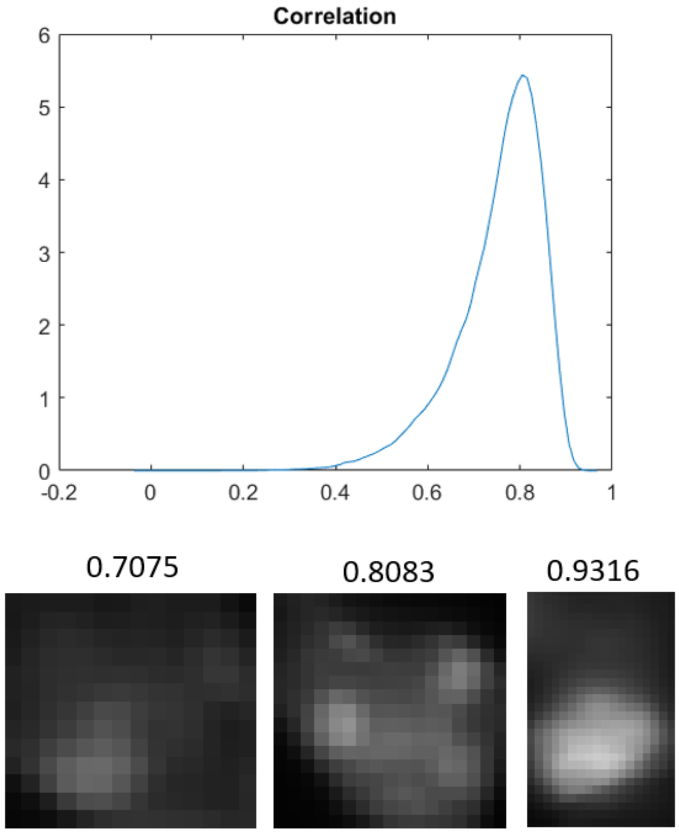


Figure 3-12: Distribution of correlation and examples of different correlation values.

Chapter 4

Tumor grade classification

The single cell features described in the previous section were used to develop models to classify tumor grade of single cells. This was done using supervised learning algorithms in which the inputs are the single cell features, and the output is the tumor grade ranging from 0 (non-tumor) to 3. I used the decision trees, random forests, adaptive boosting, and neural networks as the models. Evaluation of the algorithms was done using 10-fold cross validation.

4.1 Features

I performed a pairwise linear correlation of each feature with the tumor grade classification. I used the 14 features with an absolute value of correlation greater than 0.1 to train the models. These features are ASM, textural variance, textural entropy, Zernike amplitude (N=3, M=3), distance to nearest cell measures, HSF1 expression flag, Nkx2.1 expression flag, HSP70 log2 intensity, and HSP90 log2 intensity. The absolute value of the correlations for all the features are shown in table 4.1.

4.2 Decision tree

A decision tree is a supervised learning algorithm that classifies data by assigning data points to leaves in a tree [10]. Nodes in the tree are questions about the features, and

Table 4.1: Feature correlations

Feature	**Type**	**Correlation (abs. val.)**
Angular second moment	Texture	0.1131
Correlation	Texture	0.0939
Variance	Texture	0.2055
Entropy	Texture	0.1437
Perimeter	Morphologic	0.0337
Area	Morphologic	0.0419
Eccentricity	Morphologic	0.0860
Zernike (N=1,M=0)	Morphologic	0.0359
Zernike (N=2,M=0)	Morphologic	0.1077
Zernike (N=2,M=2)	Morphologic	0.0328
Zernike (N=3,M=1)	Morphologic	0.0034
Zernike (N=3,M=3)	Morphologic	0.0426
Cell distance (1)	Spatial	0.1781
Cell distance (10)	Spatial	0.4040
Cell distance (20)	Spatial	0.4538
Lung cell distance (1)	Spatial	0.2708
Lung cell distance (10)	Spatial	0.3789
Lung cell distance (20)	Spatial	0.3806
Ki67 +/-	Biomarker	0.0278
Nkx2.1 +/-	Biomarker	0.3197
HSF1 +/-	Biomarker	0.2696
HSP70 log2 intensity	Biomarker	0.1571
HSP90 log2 intensity	Biomarker	0.2652

given the values of the input data point, that point is assigned to a certain branch of the tree based on the point's values. This process continues from the root of the tree until the data point reaches a leaf, which represents a classification. The advantage of using decision trees is that they are usually more interpretable; however, decision trees can become very large (too many leaves and too deep), and it can be hard to decipher the contributions of individual features. I used the multiclass decision tree from Matlab's Statistics and Machine Learning Toolbox; the decision trees in this project contain over 16,000 nodes.

4.3 Random forest

Random forests are a type of bagged decision tree, which takes random subsets of the data to build decision trees and classifies new data using a majority voting scheme [18]. The data subsets are taken randomly from the original training data but with replacement, so some data points will appear more than once. Additionally, only a random subset of features are used at each node. These modifications will result in a different tree at each run. The resulting classification is obtained by combining the results of the trees and taking the most common output [10]. In this project, I used the TreeBagger from Matlab's Statistics and Machine Learning Toolbox with 10 randomly generated decision trees.

4.4 Adaptive boosting

Boosting is an algorithm that combines multiple weak decision trees into a single classifier by continuously reweighting training samples to focus on the most problematic ones. Because the individual classifiers are weak (only slightly better than random), it is typically less susceptible to overfitting. I used the Adaboost algorithm from Matlab's Statistics and Machine Learning Toolbox.

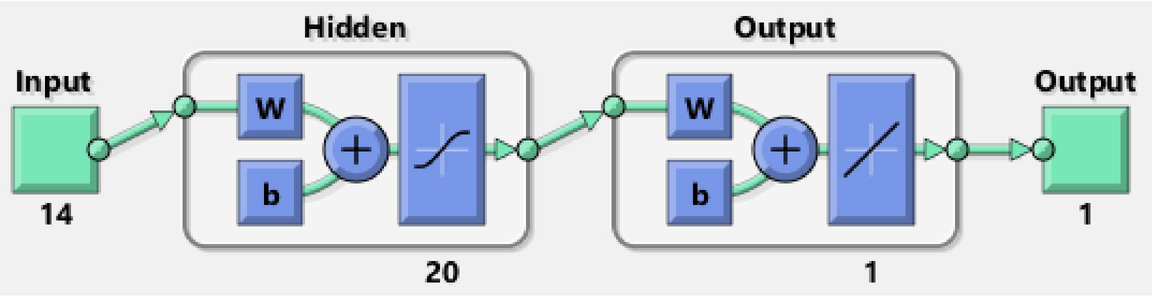


Figure 4-1: Neural net with one hidden layer

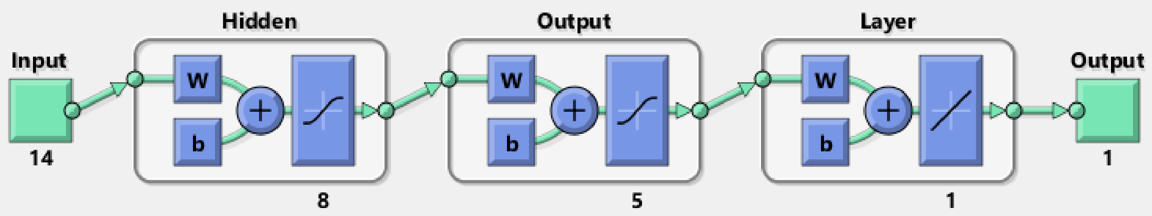


Figure 4-2: Neural net with two hidden layers

4.5 Neural network

The complexity of multi-level neural networks can potentially better capture the internal structure of large and complex datasets, but their downside is lack of interpretability [3]. In this project, I trained two feedforward neural networks from Matlab's Deep Learning Toolbox. One has 20 nodes in its one hidden layer. The other has two hidden layers, one with 8 nodes and the second with 5 nodes. The neural network setups are shown in figures 4-1 and 4-2.

4.6 Evaluation

I used 10-fold cross validation to evaluate the results. The dataset was randomly divided into 10 equal subsets. Over 10 iterations, each subset took turns being the validation data while the remaining 90% was the training data. Holding out a portion of the data when training the model guards against overfitting, while using the majority of the data as the training set guards against underfitting. I took the mode of the predictions from each iteration of the model to produce the final predictions.

Chapter 5

Discussion

5.1 Classification results

The models were trained to classify cells as belonging to non-tumor, grade 1, grade 2, or grade 3 regions. Random forests performed the best with a 72.5% accuracy. The performances of all the models are shown in table 5.1. In the initial run of the random forest model using default values from the Matlab library, there was a significant discrepancy between the performance of the algorithm on the entire dataset versus the validation set, with the validation set having an accuracy of 70% while the entire dataset had an accuracy of 99%. This indicates that the random forest model with the default parameters likely overfits for the training data and therefore will not be able to generalize well to new data. To prevent overfitting, I limited the depth of the tree, and this resulted in models that performed close to equally well on both the validation and training datasets.

The confusion matrices for the top 3 models are shown in figures 5-1, and 5-2, 5-3. Looking at the random forest matrix in figure 5-2, the sum of each row in the 4x4 table is the total number of cells that are actually in that class, while the sum of each column is the total number of cells that the model predicted to be in that class. The column summary table at the bottom of the figure shows the percentage of cells for each prediction class that match (top row) or do not match (bottom) the actual class for the cell. For the cells the model classified as non-tumor (class 0), 81.6%

Table 5.1: Tumor grade classification results

Model	**Performance (all)**	**Performance (validation)**
Decision tree	71.3%	69.6%
Random forest	72.5%	70.3%
Adaptive boosting	68.6%	68.5%
Neural network (1 layer)	56.7	56.5%
Neural network (2 layers)	59.2%	58.7%

are actually non-tumor cells, while for cells the model classified as grade 1, none are actually are grade 1 cells. The row summary table on the right side of the figure shows the percentage of cells for each class that the model classified into the correct class. For cells that are actually non-tumor, 87.5% were correctly classified as non-tumor, while for cells that are actually grade 1, only 0.1% were correctly classified as grade 1. The greatest discrepancy between the model predictions and the empirical data is in the grade 1 tumor region labeling. Only 5% of cells are in grade 1 tumor regions, and these regions tend to be smaller and more scattered throughout the tissue. Since these regions are smaller, they are also likely more susceptible to any inaccuracies in the image registration process. It is also important to note that the Aiforia model also performed significantly worse in grade 1 labeling (compared to expert labeling) than it did for the other tumor grades.

I also trained the models to predict only whether a cell belongs to a tumor or a non-tumor region. The results in table 5.2 show that all five models performed similarly well, with performances ranging from 85.2% to 87.2%. The confusion matrix of the random forest model for binary classification in figure 5-4 shows that the model correctly labeled 82.8% of the non-tumor cells and 90.0% of the tumor cells. The binary tumor versus non-tumor classification performed significantly better than the 4 category classification, possibly because the a large source of error in the 4 category classification was due to misclassifications between grades 1, 2, and 3 cells.

Out of all the features used to train the models, the spatial measures of distance between cells made the greatest impact on the model performance. Increased tumor grades correlates high cell density, which correlates with shorter distances between

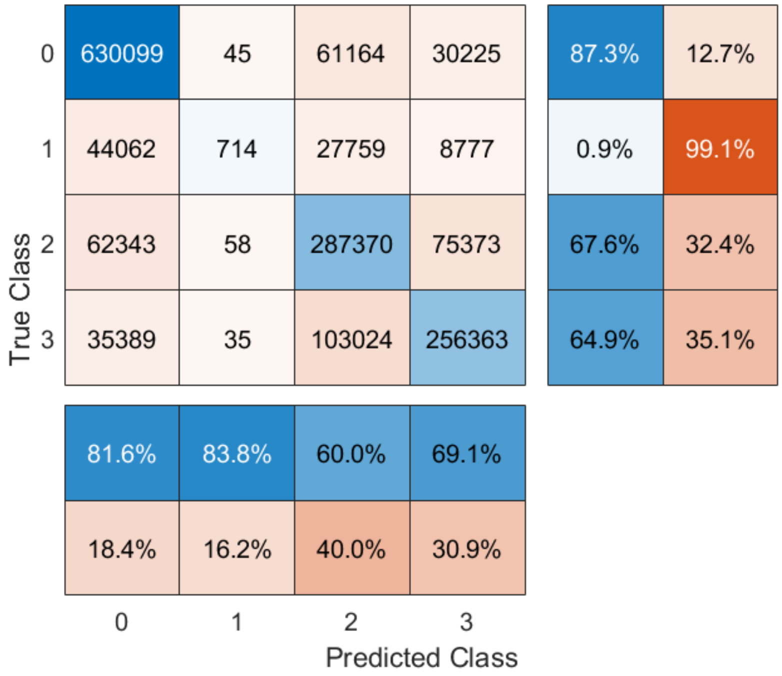
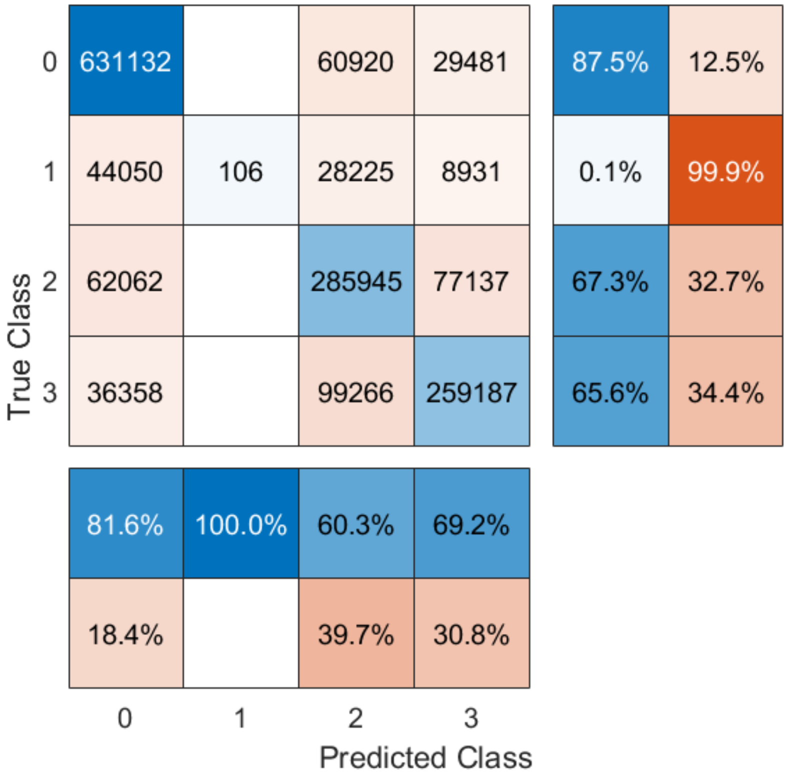


Figure 5-1: Confusion matrix for decision tree model

Table 5.2: Tumor vs non-tumor classification results

Model	**Performance (all)**	**Performance (validation)**
Decision tree	86.1%	85.1%
Random forest	87.2%	85.5%
Adaptive boosting	85.2%	85.1%
Neural network (1 layer)	85.4%	85.3%
Neural network (2 layers)	85.5%	85.4%



True Class	0	1	2	3		
0	631132		60920	29481	87.5%	12.5%
1	44050	106	28225	8931	0.1%	99.9%
2	62062		285945	77137	67.3%	32.7%
3	36358		99266	259187	65.6%	34.4%
	81.6%	100.0%	60.3%	69.2%		
	18.4%		39.7%	30.8%		

Figure 5-2: Confusion matrix for random forest model

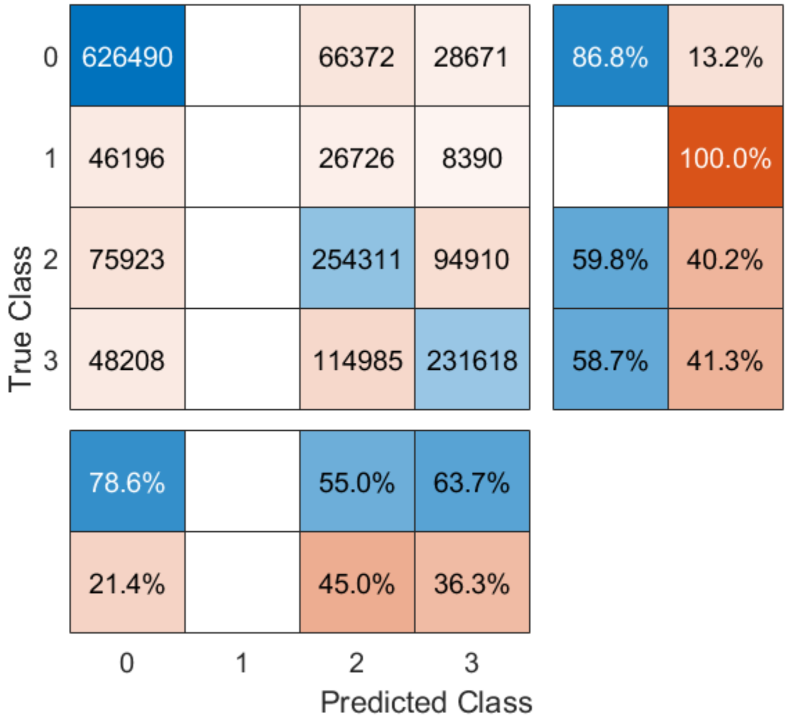


Figure 5-3: Confusion matrix for adaptive boosting model

cells. This correlation is seen in figures 5-5 and 5-6, which show boxplots of the distances between cells and their 20th nearest neighbors and 20th nearest lung cell neighbors. The top and bottom edges of the box show the 75th and 25th percentiles respectively, and the red line inside the box shows the median. The ends of the whiskers represent the non-outlier minimum and maximum in the data, and the red plus signs represent the outliers, which are data points that are more than three standard deviations away from the median.

Although previous research on histological analysis points to nuclear shape and size as factors indicative of tumor grade [8], my analysis did not find that morphological features made a significant difference in the model performance. This could be due the fact that the cell segmentation is not accurate enough to quantify small changes in nuclear size and shape. Further work needs to be done to find more informative morphological features.

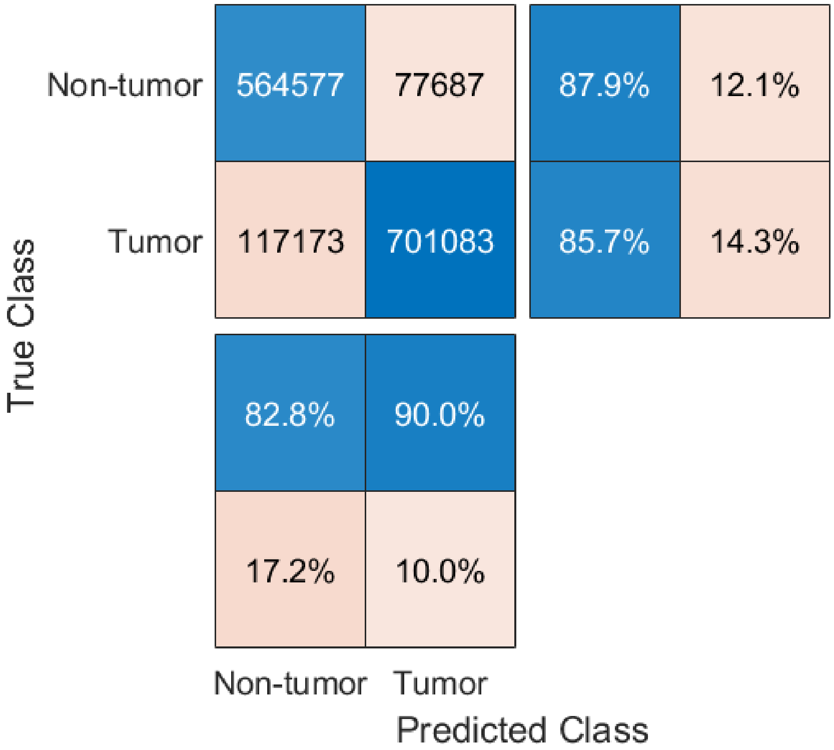


Figure 5-4: Confusion matrix for random forest model classifying tumor vs. non-tumor regions

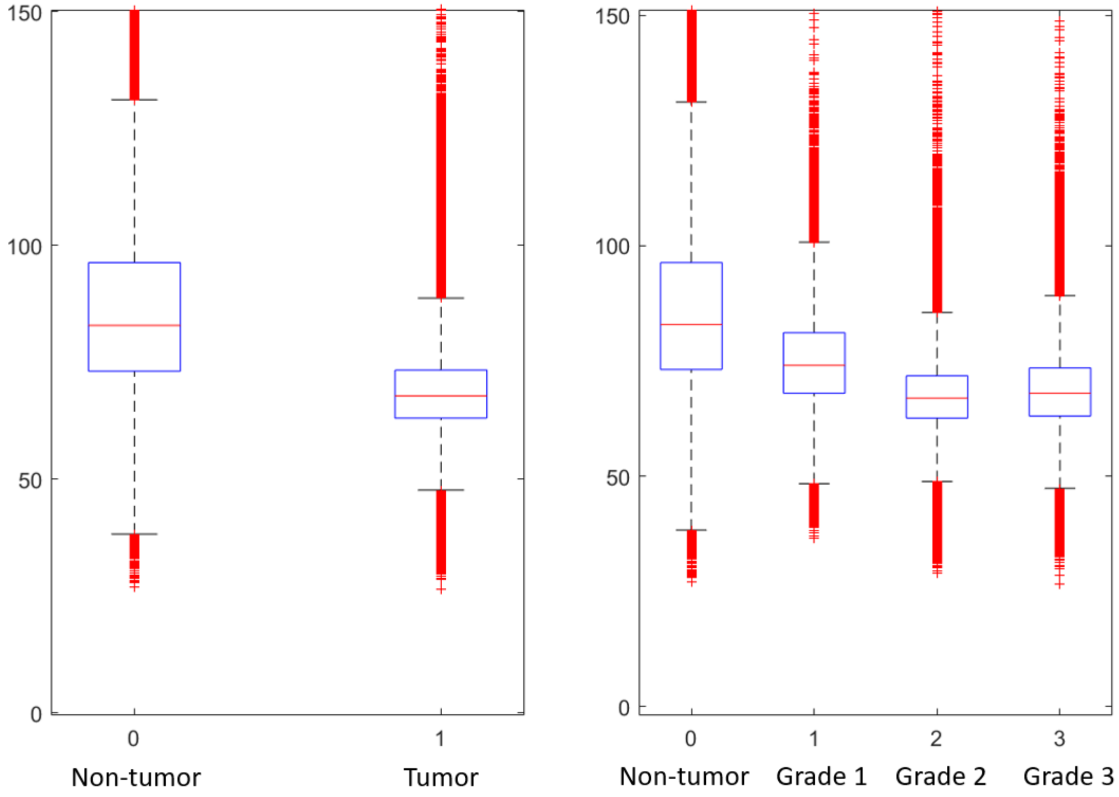


Figure 5-5: Boxplot of distance to the 20th nearest cell grouped by tumor grade.

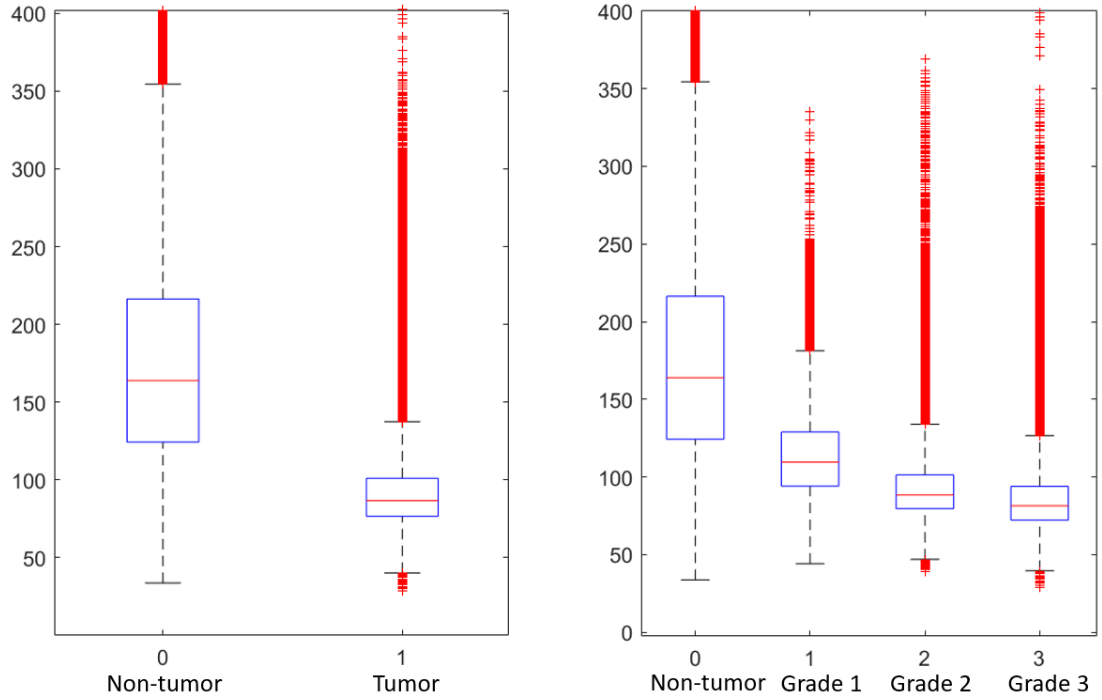


Figure 5-6: Boxplot of distance to the 20th nearest lung cell grouped by tumor grade.

5.2 Future work

This project can be expanded in the areas of data collection, data analysis, and model development. The dataset used in this project contained only 9 working biomarkers. If the number of working biomarkers can be increased, we can potentially subset cells in more biologically interesting ways. For example, certain biomarkers indicate that a cell is an immune cell, and immune infiltration of tumors is a phenomenon that could provide insight into the development of tumors.

Grade 1 tumors are the earliest legions and are typically only identified by experts through careful histological analysis. In this project, it was by far the most difficult tumor grade to correctly classify. The models classified most of the actual grade 1 tumor region cells into non-tumor or grade 2 regions. It would be interesting to investigate whether experts in the field also have difficulties classifying grade 1 tumor regions with certainty, and whether specific biomarkers or morphological features could aid in identification of these cells.

The textural features did not add significantly to the classification performance.

This could be due to the fact that the texture of the DAPI signal is not a critical determinant of tumor grade, and looking at the textures of other biomarkers could be more fruitful. For example, there is evidence that the imaging pattern of Ki67 changes as cells move into different stages of the cell cycle. Additionally, it is likely more informative to extract textural features at a larger than single cell level (e.g. crops of the image). This could address the problem of skewed textural features that arise due to badly segmented cells and could also better capture attributes of the texture that can only be detected on a more global level. For example, grade 1 and 2 tumors tend to have more uniform nuclei, while grade 3 and 4 tumors tend to have nuclei of more variable size and shape [9]. These differences may be more quantifiable when examining a larger area of cells.

Another direction of expansion is to train convolutional neural networks using cropped images as inputs. Although convolutional neural networks have the tendency to be less interpretable, they could be valuable in extracting textural and spatial information from images. Unsupervised clustering algorithms could be used cluster cells to potentially identify interesting cell sub-groups. The current models could also be further improved upon through better feature selection and hyper-parameter selection.

5.3 Conclusion

This project shows that high dimensional biomarker imaging data in addition to morphological, textural, and spatial motifs of cells can be used to predict tumor grade. Spatial features and lung cells markers in particular contribute to the model performance, which underscores the importance of using single cell imaging methods that maintain the spatial integrity of the tissue. Expanding the set of biomarkers in addition to better techniques to extract textural and morphological features could potentially improve results and lead to biologically relevant insights. The results of this project show that integrating high dimensional biomarker imaging data can be used as a tool for automated phenotyping of cells as a single cell level.

Bibliography

[1] Advancing cancer research with deep learning image analysis. https://www.aiforia.com/blog/advancing-cancer-research-with-deep-learning-image-analysis. Accessed: 2019-08-14.

[2] E. Alizadeh, S.M. Lyons, J.M. Castle, and A. Prasad. Measuring systematic changes in invasive cancer cell shape using zernike moments. *Integrative Biology*, November 2016.

[3] C. Angermueller, T. Parnamaa, L. Parts, and O. Stegle. Deep learning for computational biology. *Molecular systems biology*, 2016.

[4] E. Arvaniti and M. Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, April 2017.

[5] B. Bodenmiller. Multiplexed epitope-based tissue imaging for discovery and healthcare applications. *Cell systems*, April 2016.

[6] M.V. Boland, M.K. Markey, and R.F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, November 1998.

[7] R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, and G. P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci USA*, July 2014.

[8] M. DuPage, A. L. Dooley, and T. Jacks. Conditional mouse lung cancer models using adenoviral or lentiviral delivery of cre recombinase. *Nature protocols*, 2009.

[9] E. L. Jackson, N. Willis, K. Mercer, R. T. Bronson, D. Crowley, R. Montoya, T. Jacks, and D. A. Tuveson. Analysis of lung tumor initiation and progression using conditional expression of oncogenic k-ras. *Genes & development*, December 2001.

[10] C. Kingsford and S.L. Salzberg. What are decision trees? *Nat Biotechnol.*, June 2009.

[11] S. Klein, M. Staring, K. Murphy, M.A. Viergever, and J.P.W. Pluim. elastix: a toolbox for intensity based medical image registration. *IEEE Transactions on Medical Imaging*, January 2010.

[12] J. Lin, B. Izar, S. Wang, C. Yapp, S. Mei, P.M. Shah, S. Santagata, and P.K. Sorger. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-cycif and conventional optical microscopes. *eLife*, July 2018.

[13] M. P. Mayer and B. Bukau. Hsp70 chaperones: Cellular functions and molecular mechanism. *Cell Mol Life Sci.*, March 2005.

[14] P. Mohanaiah, P. Sathyanarayana, and L. GuruKumar. Image texture feature extraction using glcm approach. *International Journal of Scientific and Research Publications*, March 2013.

[15] I. Pantic, S. Pantic, J. Paunovic, and M. Perovic. Nuclear entropy, angular second moment, variance and texture correlation of thymus cortical and medullar lymphocytes: Grey level co-occurrence matrix analysis. *Anais da Academia Brasileira de Ciencias*, August 2013.

[16] T. Peng, K. Thorn, T. Schroeder, L. Wang, F. J. Theis, C. Marr, and N. Navab. A basic tool for background and shading correction of optical microscopy images. *Nature Communications*, June 2017.

[17] A. Prasad and E. Alizadeh. Cell form and function: Interpreting and controlling the shape of adherent cells. *Trends in Biotechnology*, April 2019.

[18] Y. Qi. Random forest for bioinformatics. *Ensemble machine learning*, 2012.

[19] F. Saki, A. Tahmasbi, H. Soltanian-Zadeh, and S.B. Shokouhi. Fast opposite weight learning rules with application in breast cancer diagnosis. *Comput. Biol. Med.*, 43(1):32–41, 2013.

[20] F.J. Sanchez-Rivera, T. Papagiannakopoulos, R. Romero, T. Tammela, M.R. Bauer, A. Bhutkar, N.S. Joshi, L. Subbaraj, R.T. Bronson, W. Xue, and Jacks T. Rapid modelling of cooperating genetic events in cancer through somatic genome editing. *Nature*, December 2014.

[21] F.H. Schopt, M.M. Biebl, and J. Buchner. The hsp90 chaperone machinery. *Nature Reviews Molecular Cell Biology*, April 2017.

[22] C. Sommer, C. Straehle, U. Kothe, and F. A. Hamprecht. Ilastik: Interactive learning and segmentation toolkit (2nd. edition). *Eighth IEEE International Symposium on Biomedical Imaging (ISBI)*, January 2011.

[23] A. Tahmasbi, F. Saki, and S.B. Shokouhi. Classification of benign and malignant masses based on zernike moments. *Comput. Biol. Med.*, 41(8):726–735, 2011.

[24] A. Vihervaara and L. Sistonen. Hsf1 at a glance. *Journal of Cell Science*, 2014.