Automated Image Analysis for Nuclear Morphometry Using H&E and Feulgen Stains in Prostate Biopsies

BY

KUSUMA BAPURE
B.E., Visveswaraya Technological University, 2009

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Chicago, 2011

Chicago, Illinois

Defense Committee:

Rashid Ansari, Chair and Advisor
Vldamir Goncharoff
Peter Gann, Pathology

To my Parents

ACKNOWLEDGEMENTS

I would like offer my gratitude to all individuals who were with me every step of the way to my thesis. It is difficult to overstate my gratitude to my supervisor, Dr. Rashid Ansari, for his inspiration and enthusiasm to supervise my work. I am extremely grateful to my thesis committee, Dr. Peter Gann, Dr. Yi Lu and Ryan Deaton for all their guidance and help during my thesis research and writing. I sincerely appreciate my team members – Dr. Gann, Dr. Yi Lu, Ryan Deaton, Dr. Anup Ametya and Peter Nguyen for the support and encouragement during the time spent at the Pathology Research laboratory. My sincere regards to Dr. Vldamir Goncharoff, who as a committee member provided valuable insights.

My special thanks are due to Ryan for his great efforts to explain things in a clear and simple manner, and for his help to make biology fun for me. Throughout my thesis-writing period, he provided continuous encouragement, sound advice, good teaching, good company, and lots of good ideas. I would have been lost without the wonderful work environment he created and for his constant support.
I would like to thank the staff of the Department of Electrical and Computer Engineering who helped me along the way, in particular Agustina Alvarado and Alicja Wroblewski who took the time to kindly answer numerous questions and guide me through some difficult paper work.

My deepest gratitude goes to my parents, especially my mother, for all her sacrifice and constant support, to my genius scientist and technocrat uncle and to my extended family of friends, comprising of Mahesh, Nischit, Venkat, Surabhi, Deepak, Vaidehi and many others, who have stood by me during this exciting journey and testing times. This would not have been as much a serious job and fun, without their warmth and special care. Finally, I would like to also thank Dr. Gann, Dr. Richard Magin and Dr. John R. Hetling for giving me the opportunity to work in the lab, extending the all

important financial support during my thesis project, and more so, for letting me be a part of this wonderful family.

KB

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

PCa	Prostate Cancer
EM	Expectation Maximization
PIN	Prostate Intraepithelial Neoplasia
HGPIN	High-Grade Prostate Intraepithelial Neoplasia
DNA	Deoxyribonucleic acid
PSA	Prostate Specific Antigen
RSV	Radial Symmetry Voting
QNG	Quantitative Nuclear Grade
AOD	Average Optical Density
AUC	Area Under Curve
ROC	Receiver Operating Characteristics
StdOD	Standard Optical Density
LR	Logistic Regression
MFS	Multi Feature Score
nMFS	Nuclear level MFS
pMFS	Person level MFS

SUMMARY

The thesis addresses the problem of analyzing prostate cancer biopsy image data. It describes an unsupervised semi-automated method for segmentation of nuclei in Hematoxylin and Eosin (H&E) stained Prostate Biopsy images and investigates its use in modeling the data and its effectiveness in predicting cell alterations. Existing methods have largely focused on the use of Feulgen-stained prostate cancer biopsies in the analysis of Nuclear Morphometry due to its DNA staining capability. In this thesis the potential for the use of the more easily available H&E staining is investigated since H&E stains are widely used in medical diagnosis due to the simplicity in the staining procedure. Our results provide evidence that the H&E can yield a performance comparable to a reference method that uses Feulgen-stained biopsies.

Over the years the growth of digital pathology has increased due to its advantages over manual segmentation techniques. The use of clinical trials to obtain sufficient patient data for the prevention of cancer requires time consuming and laborious procedures. However, measuring histological sections of biopsies and smears might reveal positive cancer prevention methods within lesser time limits and fewer subjects. We are currently in search of new molecular biomarkers to predict the biochemical recurrences and alterations at the cell and molecular level in histologically normal looking tissue. Nuclear Morphometry has been a relatively new approach to assess the pre - diagnosis of cancer via digitized histology. A set of over 180 features are extracted from each image in a database of 42 H&E stained negative prostate biopsies. These are grouped into Cases (subjects with no cancer on their initial biopsy and subsequently received a cancer diagnosis) and Controls (followed for an equivalent period of time without cancer being detected). However, until now, only Feulgen has been used in the analysis of Nuclear Morphometry due to its DNA staining capability. In this study, prostate biopsies stained with H&E are studied as test and training samples and the performance are compared with the Feulgen stained slide results.

A robust algorithm (implemented in MatLab) has been developed for semi-automated segmentation of glandular structures in histopathology imagery (Aperio ScanScope). K-Means clustering algorithm and other morphological operations are used to pre-process the images to filter out irrelevant structures. The nuclei centers obtained with Radial Symmetry transform act as markers for marker controlled watershed segmentation. The approach detects multiple nuclei from a closely spaced/merged cluster of nuclei. Architectural and texture features were measured for each cell image. The method provides good performance in terms of segmentation accuracy. In this preliminary study, the good agreement between the morphometric results and general histomorphologic data demonstrates the importance of nuclear morphometric analysis using H&E stains in benign prostate biopsies, which could be extended to other cancer types. The person-level Multi-Feature Score (pMFS) is produced by applying Logistic Regression to the reduced feature set. Its result is verified by an Area-Under-receiver-operating-Curve (AUC) value of 0.77.

1. INTRODUCTION

1.1. Overview

Computer Aided Diagnosis (CAD) has made remarkable progress in the field of medicine by helping pathologists diagnose cancer recurrences with improved accuracy, reproducibility and efficiency. Cell nuclei segmentation is an important step towards automatic analysis of digitized histological slides of prostate cancer biopsies. Existing methods have largely focused on the use of Feulgen-stained prostate cancer biopsies in the analysis of Nuclear Morphometry (M Guillaud, p.34, 2005) due to its DNA staining capability. In this thesis the potential for the use of the more easily available Hematoxylin and Eosin (H&E) staining is investigated since H&E stains are widely used in medical diagnosis due to the simplicity in the staining procedure. Our results provide evidence that the H&E can yield a performance comparable to a reference method that uses Feulgen-stained biopsies.

Manual outlining of the cell nuclei has not only proved to be labor intensive but also has suffered from observer variability with regard to the irregularity in contour characteristics(H. Fatakdawala, p.1676, 2010) of cancer cells. The diversity and complex structure of the tissue architecture makes the nuclei segmentation a complex task. The use of automatic analysis of histological slides has significantly increased over the years, especially in prostate cancer diagnosis. Many cell segmentation algorithms have been developed over the past years which reject the malformed outlines using Support Vector Machine (SVM). Watershed segmentation, Expectation Maximization (EM) (A Hafiane, p.903, 2008) and other standard methods used for nuclei detection work well only with

considerable separation between the nuclei. Hence, there is a need to develop a good segmentation technique that can identify individual nuclei obtained with different stains and classify them based on morphological features (texture, shape, size etc.). We believe that field effects have not been studied before extensively using nuclear morphometry and have a large potential for use in epidemiological research.

1.2. Main Contributions

The major objective here is to interpret the changes in shape, size and chromatin texture (Malik et al., p.7, 2001) of nuclei by digital morphological morphometric analysis and to build a statistical model which can differentiate changes associated with Cases and Controls. We also aim to develop and validate a criterion in the form of a Multi-Feature Score (MFS) based on DNA staining that accurately discriminates case and control population, ultimately allowing Hematoxylin and Eosin (H&E) to be used as a stain to detect field effects in benign tissues.

The approach detects multiple nuclei from a closely spaced/merged cluster of nuclei. Architectural and texture features are measured for each cell image. In this preliminary study, the good agreement between the morphometric results and general histomorphologic data demonstrates the importance of nuclear morphometric analysis using H&E stains in benign prostate biopsies, which could be extended to other cancer types.

The cell segmentation procedure developed in this thesis differs significantly from the widely used watershed segmentation method (L Latson, p.321, 2003). In comparison with the Feulgen-stained samples the H&E samples are more challenging with respect to segmenting the overlapping objects. A suitable automated technique is required for successfully segmenting the high resolution images of the epithelial nuclei. Hence a Marker Controlled Watershed Segmentation method is adopted here wherein the regional minima used as markers are obtained by Radial Symmetry Voting based technique (A Hafiane, p.903, 2008). K-Means algorithm generates the background markers and the Radial Symmetry technique (Q Yang, p.63, 2004) generates the foreground markers. Then, by applying watershed segmentation, one is successfully able to identify the required epithelial nuclei. Morphological filters are used to remove the false positives detected (stromal cells, RBC). A 2-step model pMFS (person level Multi-Feature Score) obtained after a dimension reduction of the feature set is shown to yield an Area-Under-Curve (AUC) value of 0.77 for the selective feature set (obtained by backwards elimination) as a comparison of tissues with different histologic types stratified by their malignancy associated changes. Previously, the Prostate Biopsies stained in Feulgen were segmented using K-Means and Watershed segmentation alone and yielded an AUC of 0.76.

1.3. Organization

The remainder of the thesis is organized as follows: Chapter 2 describes the background for research, computer-aided diagnosis (O Sertel, p.2613, 2010) for prostate cancer prediction and the role of Nuclear Morphometry as biomarkers. In Chapter 3, image acquisition and analysis, including staining procedures, image acquisition, and

image segmentation techniques, are discussed. the feature data prediction model is examined in Chapter 4. Results obtained from statistical analysis conducted on Feulgen versus H&E and Case/Control prostate biopsies are presented in Chapter 5. Finally, the research conclusions and some topics for future studies are highlighted in Chapter 6.

2. BACKGROUND

2.1. Prostate Gland

The human prostate gland is a walnut-sized organ homologous to men. It is situated at the base of the bladder, between the pubic bone and the rectum surrounding the urethra[1]. The gland secretes an alkaline, white fluid constituting about 25% of the seminal fluid which combines with the sperm, during ejaculation. The prostatic fluid acts as a lubricant to prevent infection in the urethra protecting and energizing the sperms.

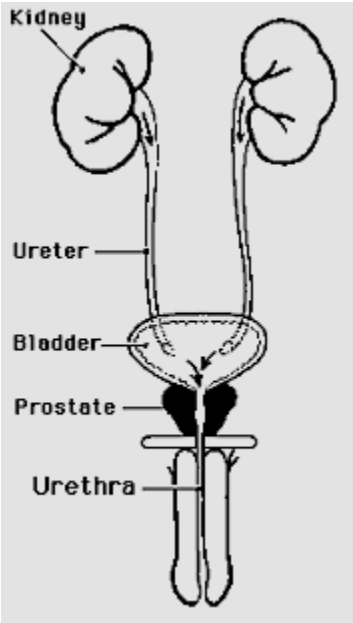

Figure 1 - Anatomy of the Prostate Gland

(Source: National Kidney & Urologic Diseases Information Clearinghouse)

2.2. Functions of the Prostate Gland

The alkalinization of semen is primarily accomplished through secretion from the seminal vesicles. The prostatic fluid is expelled in the first ejaculated fractions, together

[1] the tube that carries urine from the bladder to the penis

with most of the spermatozoa. In comparison with the few spermatozoa expelled together with mainly seminal vesicular fluid, those expelled in prostatic fluid have better motility, longer survival and better protection of the genetic material (DNA).

2.3. Prostate Cancer (PCa)

Other than lung cancer, cancer of the prostate is the most common cancer among American men. According to the latest estimates for prostate cancer in the United States (American Cancer Society 2011), there are about 240,890 new cases of prostate cancer diagnosed and 33,720 (approx.) men will die of prostate cancer annually. One in every 6 men will be diagnosed with prostate cancer during his lifetime. However, more than 2 million men in the United States who have been diagnosed with prostate cancer at some point are still alive today. At the same time, about 1 man in every 36 will die of prostate cancer.

Determining the stage at which prostate cancer inflicts a patient is essential in the diagnosis of cancer. Prostate cancer is determined by the PSA (Prostate Specific Antigen) Diagnosis is done through a biopsy where a small section of the prostate gland (less than a percent of the gland) is sampled for examination. In molecular level, the rapid growth of malignant cells increases with the growth of the disease and can spread to other parts of the body. Early detection of the disease would result in a very high chance of cure. Men undergo at least three or four biopsies before cancer is detected. It is natural for men to worry about having prostate cancer despite negative results from an initial biopsy. Researchers have included seven independent risk factors for a positive repeat prostate biopsy after an initial false-negative biopsy. These include:

1. Age of patient

2. Family history of PCa
3. PSA density
4. Abnormal prostate exam results
5. Length of time taken for the PSA level to double
6. Number of samples taken during the initial biopsy, and
7. Presence of abnormal cells (HGPIN - High-Grade Prostatic Intraepithelial Neoplasia)

Patients with high risk features mentioned above continue to remain at risk of being detected with PCa even with negative biopsies. The figure below shows the progression of the epithelial cells. Dysplasia refers to the abnormal development of immature cells. The disruption of the basal cell layer occurs in response to the luminal carcinogens and seems to be a prerequisite to the stromal invasion.

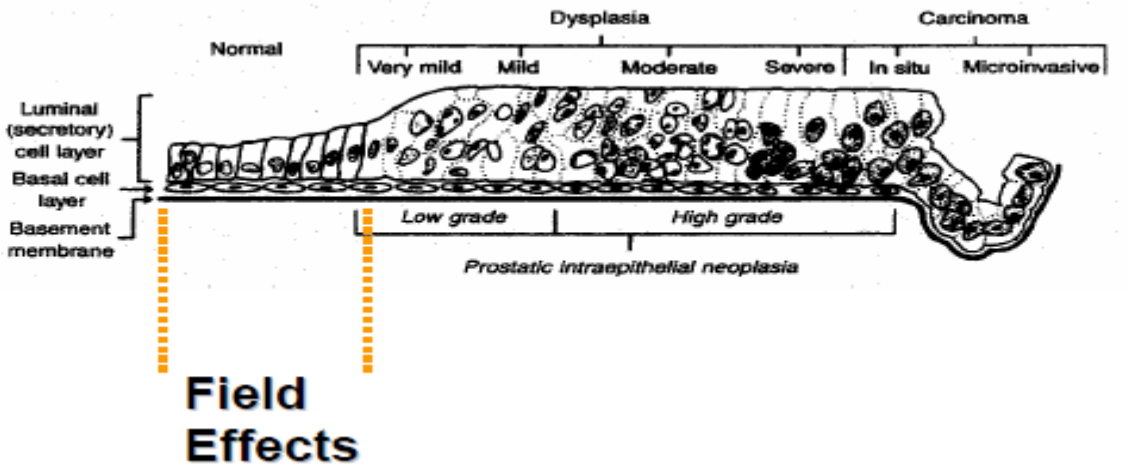

Figure 2 - Tissue progression from a normal prostate to cancer[2]

[2] McNeal JE, et al. *Prostate* 27:258-68, 1992

2.4. Computer Aided Diagnosis of Digital Pathology

Conventional pathological techniques involve increasing risk for human error in slide preparations and patient identification. The recent advent of cost effective whole slide scanners has enabled the digital transformation of patient and laboratory medicine. The development of image analysis software has increased the potential to provide better methods for quality assurance and image comparison. More importantly, it has opened a plethora of slide-based technologies to pathologists. Digital image analysis has made remarkable progress in cancer prediction and prevention. Studying the morphometry can provide the pathologists with valuable information about the course of treatment and determine the relative effect of drug candidates in cells (cytology) and tissues (histology). The main challenges in analysis include the enormous density of data. For example, a single prostate biopsy tissue sample when digitized consists of approximately 225 million pixels and each of the prostate biopsy procedure comprising up to 4 billion pixels (A. Doudkine, p.286, 1995).

Along with digitization, identifying prognostic markers and prediction of the disease recurrences plays an essential role in therapy. The role of a pathologist remains constant throughout the procedure of quantitative cross modal data integration for diagnostic and prognostic purposes.

2.5. Role of Nuclear Morphometry as Biomarkers

Biomarkers, also called as clinical indicators, play a very important role in risk identification and modulation by chemo preventive agents. Biomarkers that can estimate the intensity of treatment and determine the likely outcome of the disease are termed as

prognostic biomarkers. They are capable of distinguishing between those patients who should avoid over treatment of an indolent form and those who should be treated properly to stop the aggressive form of the disease. By identifying the cell population the markers can indicate the progress and effects of the disease and this creates a potential for monitoring the early events in cancer evolution. In case of prostate cancer, the most commonly used marker is PSA which detects the increased production in serum, which generates false-positive results, since it is incapable of differentiating between the cancerous tissues, (Benign Prostatic Hyperplasia) BPH and (Prostatic Intraepithelial Neoplasia) PIN. Hence, an adequate molecular marker is essential for early diagnosis and progression.

Specially designed cameras, such as those based on Charge-Coupled Devices (CCD), are capable of capturing high-resolution images with varying parameters to process the nuclei characteristics. If specific DNA stains are used, the chromatin distribution in the cell nuclei can be measured which plays a crucial parameter in cytopathology. In tumors, the chromatin changes are seen to be associated with the progression of disease. Nuclear Morphometry covers a wide range of nuclei features including shape, size, intensity, texture, run length, fractal, texture and Markovian texture descriptors used to predict patient outcome in PCa.

Since 1982, nuclear morphometry has been used as a predictive measure in the prognosis of PCa. Nuclear morphometric alterations measured by computer aided image analysis is able to detect abnormal DNA content representing large scale chromosomal alterations which reflects genetic instability in tumor cells. Various studies have been developed to give quantitative results for cancer prognosis. Eichenber and associates

developed 12 features to predict the ellipticity factors, nuclear roundness, and other concavity features. In a similar study done by (Partin A. W. et al., p.1254, 1989) 15 descriptor features were calculated to compare the Gleason grading as a predictor of prognosis of A2 prostate cancer.

Table I – 15 Basic Features used in calculating the Gleason Grade Study

1	Area	9	Sub-Optimal circle fit
2	Perimeter	10	Variance
3	Form Factor	11	Standard Deviation
4	Roundness	12	Minimum Diameter
5	Ellipticity (Feret)	13	Maximum Diameter
6	Ellipticity (Inertia)	14	Sum of Squares
7	Bending Energy	15	Range
8	Convexity		

Quantitative Nuclear Morphometry (QNM) along with the spatial and textural analysis helps generate the digital image cytometry (A Doudkine et al., p.286, 1995).

2.6. Image Cytometry Significance

1. Preservation of the architecture of the tissue under study.
2. Histological information is used to select and classify the cells.

Along with the shape and size of the cells, chromatin material affects the pattern of genes' activation. Thus, the morphological features calculated are used for cancer diagnosis (Pamela Wolfe, p.976, 2004).

Table II - Size/Shape and Pixel/Texture dependent features used in Nuclear Morphometry

Features	Category	Description
Nuclear Area	Size	Total no. of pixels covered by the nucleus
Perimeter	Shape	Total no. of pixels bordering the nucleus
Circularity	Shape	Square of the perimeter divided by the area. Gives the measure of the roundness of the nucleus. Compares the perimeter of the nucleus to the total size of the nucleus
Summed optical density	DNA Content	Sum of each individual intensity value over all pixels comprising the nuclear body
Average optical density	DNA Content	Summed optical density divided by the number of pixels comprising the nuclear body
Cell Feret X	Shape	Width of a bounding rectangular box around the nucleus (short side)
Cell Feret Y	Shape	Height of a bounding rectangular box around the nucleus (long side)
Maximum Diameter	Shape	Maximum diameter of the nucleus(through centroid of the nuclear)
Minimum Diameter	Shape	Minimum diameter of the nucleus(through centroid of the nuclear)
Elongation	Shape	Ratio of maximum diameter by minimum diameter. Highly circular nucleus has a value of 1
Coarseness	General Texture	Count of slope –

		(2*count of peak – count of valley)
Valley	General Texture	Both of the neighbor pixels have gray level values higher than the currently evaluated pixel
Slope	General Texture	Both of the neighbor pixels have gray level values less than the currently evaluated pixel
Peak	General Texture	One of the neighbor pixels gray level value is less than the currently evaluated pixel, and one of the neighbor pixels gray level is greater
Margination Range in [0 1]	DNA Content	Reflects DNA Mass distribution along radial direction 0 : Mass concentration at the center of the nucleus 1: Entire mass distributed along the nuclear border

There were 17 different statistical analysis tests done which included distribution tests, variability tests, and extreme tests. It was observed that nuclear roundness alone was one of the important descriptor and provided a good separation of subjects by the prognosis.

The figure below represents the valley, peak and slope features where the colors are represented as:

Black - Highest pixel intensity,

White - Lowest pixel intensity and

Gray – Level between black and white

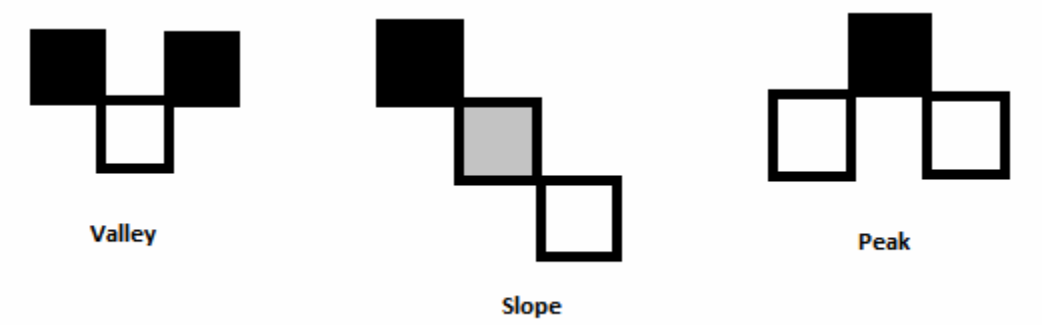

Figure 3- Threshold Features with count detection

1. Photometric Features:

They describe the chromatin texture of the nuclei. Typically for meaningful measurements, the nucleus of a cell occupies 200- 450 pixels.

OD Kurtosis: It is a measure of the vertical distribution of the optical density and it is equal to 3.0 for normal distribution.

2. Discrete Texture Features:

Discrete texture features are unique features which depend on the threshold of OD distribution of the nuclei. They are very sensitive to the focus set while generating snapshots by the microscope. The Optical Density (OD) values are calculated per pixel in the image of the nucleus. Each pixel in the 100:1 has a range between 255 (black) and 0 (white) and the overall OD is calculated by the sum of all OD values of all pixels in the defined image. Two thresholds are set with reference to the Average OD of the normal epithelial nucleus of the population and different features are calculated using these thresholds.

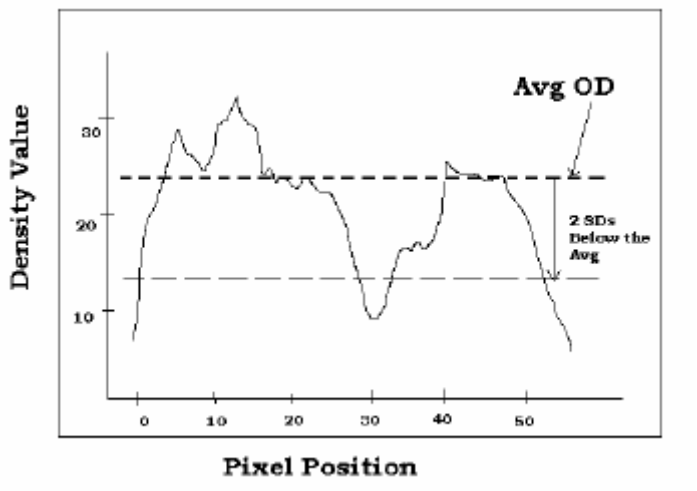

Figure 4 - The Image texture distribution shows density value plotted against pixel position. The lower threshold (LT) and upper threshold (UT) are decided using the features average optical density (AOD) and standard deviation (SD) values of the nuclei

Standard Deviation: This feature is computed as the standard deviation of the optical density of every pixel that composes the cell.

Area: Represents the total nuclear area occupied by low/medium/high chromatin condensation state (Low DNA Area, Medium DNA Area, and High DNA Area).

DNA Amount: The ratio of integrated optical density of the low/medium/high condensation state to the total integrated optical density **(**Low DNA Amount, Medium DNA Amount, High DNA Amount).

Object Density: Count of low/medium/high density regions inside the nucleus

DNA Compactness: Measure of the shape of the low/medium/high chromatin state components.

Center Mass: Measure of low/medium/high asymmetry distribution of chromatin regions

Fractal Dimension Features: Measurement of the rate at which the fractal area increases at finer scales of the area of a 3D surface created by the 3D plot of OD and x, y co-ordinates.

These features are developed using thresholds on the texture area drawn as shown in Figure 3.

3. Blob and Hole Features:

To try the different thresholds within the distribution various iterations were done by changing the thresholds with different combinations of calculation. A feature vector which included different threshold features was calculated for 8 different iterations resulting in different upper threshold and lower threshold values.

Combination 1: Upper threshold = AOD, Lower threshold = AOD – (1 x SD)

Combination 2: Upper threshold = AOD, Lower threshold = AOD – (1.5 x SD)

Combination 3: Upper threshold = AOD, Lower threshold = AOD – (2 x SD)

Combination 4: Upper threshold = AOD, Lower threshold = AOD – (2.5 x SD)

Combination 5: Upper threshold = (AOD + SD), Lower threshold = AOD – (1 x SD)

Combination 6: Upper threshold = (AOD + SD), Lower threshold = AOD – (1.5 x SD)

Combination 7: Upper threshold = (AOD + SD), Lower threshold = AOD – (2 x SD)

Combination 8: Upper threshold = (AOD + SD), Lower threshold = AOD – (2.5 x SD)

4. Markovian Features:

These unlimited set of histogram probability dependent features are calculated using long calculations taking into consideration small region of the nuclei at a time. It

describes the variations among the adjacent pixels within the nucleus. The initial step begins with a small area of 9x9 pixels and extends to cover the entire nucleus area calculating various features like difference-mean, shade, contrast, correlation and 2nd angular moment (Bacus et al., 1987) The human perception is very similar to the distance in the perceptually uniform color space. Also, the L*a*b* space is a measure of the perceptual distance marked by the average human in the Euclidean distance. Hence, we make use of the L*a*b* distribution for the color distances.

Entropy: It is defined as the disorder in the measure of the sum and difference histograms, wherein large values correspond to much disorganized histograms. It characterizes the texture of the image.

Contrast: The measure of the contrast is based upon the differences between neighboring pixels, thus higher value of contrast indicates large variations in the pixels optical density.

Homogeneity: It is opposite to contrast and measures the smoothness in the image intensity. Large value indicates more structural uniformity.

Cluster Shade: The Markovian texture gives large values for images with a few distinct chromatin clumps, negative values are dark clumps on light background and positive values are light clumps on dark background.

Cluster Prominence: Large values indicate the predominance of the chromatin clumps to have a higher contrast to the background.

Table III - Texture Based Features on the Intensity Histogram of a Region

Mean: $m = \sum_{i=0}^{L-1} z_i p(z_i)$	Measure of average intensity z_i is a variable indicating intensity $P(z)$ is the histogram of the intensity levels in a region, L is the number of possible intensity levels (for example:256)
Standard deviation: $\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2}$	Measure of average contrast. The larger the standard deviation, the coarser
Smoothness: $R = 1 - 1/(1+\sigma^2)$	Measures the relative smoothness of the intensity in a nuclear. R is 0 for a constant intensity (smooth), 1 for regions with large excursions in the values of its intensity levels
Third moment (symmetry): $\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$	Measure the skewness of a histogram. This measure is 0 for symmetric histograms, positively by histograms skewed to the right (about the mean) and negative for histograms skewed to the left
Uniformity $U = \sum_{i=0}^{L-1} p^2(z_i)$	Measure is maximum when all gray levels are equal and is directly proportional to smoothness
Entropy $e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$	Measure of randomness. Entropy is directly proportional to the random intensity distribution

2.7. Conceptual Framework of the Study

We believe that benign prostate biopsies yield important benefits in both clinical practice and chemoprevention. However, measuring histological sections of biopsies and smears might reveal positive cancer prevention methods within lesser time limits and fewer subjects. We are currently in search of new biomarkers to predict the bio-chemical recurrences and alterations at the cell and molecular level in histologically normal looking tissue. Nuclear Morphometry has been a relatively new approach to assess the pre - diagnosis of cancer via digitized histology. A set of over 180 features are extracted

from each image in a database of 42 Hematoxylin and Eosin (H&E) stained negative prostate biopsies. Grouped as Cases (subjects with no cancer on their initial biopsy and subsequently received a cancer diagnosis) and Controls (followed for an equivalent period of time without cancer being detected). H&E stains are widely used in medical diagnosis due to their simplicity in staining and availability. However, until now, only Feulgen has been used in the analysis of Nuclear Morphometry due to its DNA staining capability. In this study, prostate biopsies stained with H&E are studied as test and training samples in comparison with the Feulgen stained slide results.

A high precision algorithm (implemented in MatLab) has been developed for automatic image segmentation of tissue components, glands, and epithelial cell nucleoli. The glandular structure identification in histopathology imagery was carried out using the digital pathology system-Aperio ScanScope). Using a unique combination of color (using K-Means) and object (Radial Symmetry and Watershed) an accurate segmentation technique was developed to overlook the false positives. The approach detects multiple nuclei from a closely spaced/merged cluster of nuclei. Architectural and texture features were measured for each cell image. In this preliminary study, the relative concordance of the morphometric results and general histomorphologic data exhibited the importance of nuclear morphometric analysis in benign prostate biopsies, which could be extended to other cancer types. A 2-step model pMFS (person level Multi-Feature Score) is computed in SAS and generates an AUC of 0.77 for the selective feature set (obtained by backwards selection) as a comparison of tissues with different histological types stratified by their malignancy associated changes.

3. IMAGE ACQUISITION AND ANALYSIS

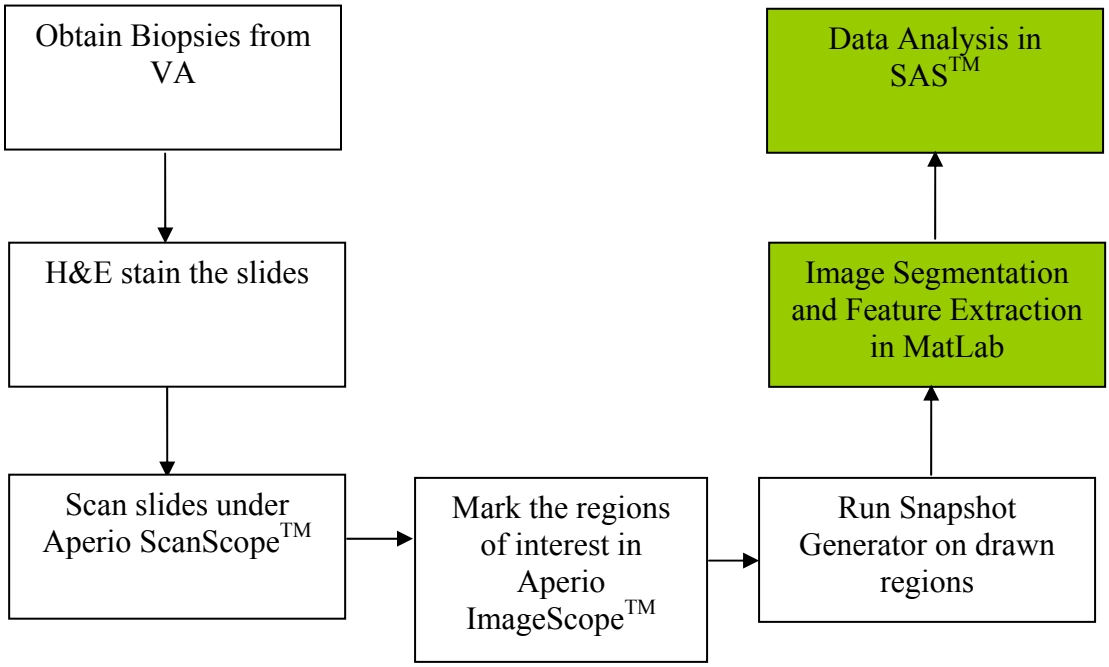

Figure 5 - Flow Diagram of the steps involved in the study. My contribution involves the green boxes (Image Segmentation, Feature Extraction and Data Analysis).

3.1. Biopsies

The prostate biopsies used in the study were split into two groups:

1. Cases

All case subjects must have biopsy-confirmed prostate cancer (PCa) and an earlier negative biopsy made at either the Jesse Brown VA Medical Center or the Lakeside VA. The earlier biopsy should be performed in 1997 or later and made sure not to be fixed in Bouin's and has had no treatment with finasteride prior to index biopsy.

2. Controls

Controls must match to the case on age within three years. Negative biopsy must

match date of index biopsy within 90 days and have no record of diagnosis of PCa. They must have two negative biopsies in addition to index biopsy, which can be retrieved before or after index biopsy. Prostate specific antigen must be less than 10 mg/mL throughout their record. Biopsies are required neither to be fixed in Bouin's nor be treated with finasteride prior to index biopsy.

3.2. Staining

While considering intermolecular effects dye uptake is often thought of as involving bonding types of strong, directed attractions. If the tissue component has higher affinity for dye unlike its surroundings then we are likely to see staining. Staining also depends upon the rate of staining effects. Progressive staining is when a tissue component takes up the dye more rapidly than its surroundings. Regressive staining occurs if the target loses dyes more slowly than its surroundings and if the destining is terminated before losing out on most of the dye. Morphology and stain protocols have a close association due to their capability of showing the texture, shape/size features of the glandular structures studied.

3.2.1. DNA Ploidy

Normal tissues and neoplasm are considered to be cycling if the DNA is synthesized actively during cell division. Cell proliferation is a continuous process, which is ongoing in a normal tissue by which process the dead cells are replaced with new ones. A cell which is not actively proliferating is considered in a diploid state. The DNA content of a cell in diploid state contains a normal number of chromosomes, which increases with the proliferation of the cell. The analysis is done to assess the

characteristics of prostate cancer cells usually done after a prostate biopsy and is capable of measuring their number.

3.2.2. Feulgen Stain

Feulgen is a special stain which binds stoichiometrically with the double stranded DNA molecule. (Mariuzzi et al p.87, 2000) studied the importance of Feulgen staining because they realized that the main changes in tumors are the cellular changes in the nucleus. (Veltri et al ., p.102, 2008) stated that normal physical changes such as cell division and disease responses such as cancer can change the nuclear structure. Thus, measuring nuclear structure precisely can give great benefits when diagnosing a disease. Since chromatin is mostly concentrated in the nucleus, Mariuzzi studied the nuclear distribution and characterized them in the difference in lesions. Mariuzzi and Veltri are one of the many who studied nuclear morphometry with Feulgen due to the stain's specificity of detecting chromatin. The Summation of the Optical Density of each Feulgen stained nucleus can be used to calculate the amount of the DNA present based on the Beer-Lambert's Law.

3.2.3. Protocol for Feulgen Staining[3]

1. Hydrate slides to water (dry smears should be placed in water for 2 to 5 minutes).
2. Hydrolyze in 5N HCl for 60 minutes at room temperature (15°C to 30°C). For quantitative work, the hydrolysis time is critical. At this temperature and acid strength, the extended exposure time places fewer demands on precise timing of acid

[3] Image Path Systems Feulgen Staining Protocol 2008

exposure. With all methods of hydrolysis, a hydrolysis curve should be generated to verify that an appropriate protocol is being used[4].

3. Rinse slides in deionized water to remove excess acid.
4. Stain in the decolorized ImagePath Blue Feulgen Stain solution for 60 minutes at room temperature. The staining dish should be covered, but does not have to be sealed. A narrow blue band may appear on the top of the stain solution during staining, but this will not affect staining[5].
5. Rinse in running water or three changes of deionized/distilled water for a total of five minutes.
6. Place slides into three separate changes of ImagePath Rinse Reagent, 5 minutes each, for a total of 15 minutes. Rinse solution should be covered during use, to slow liberation of SO_2 into the laboratory.
7. Rinse slides in running deionized water, or use three separate changes of deionized/distilled water.
8. Place slides in acid alcohol for 5 minutes.
9. Dehydrate slides, starting with 70 % ethanol (fresh, previously unused) for 3 minutes.
10. Continue dehydration in two changes of 100% ethanol (fresh, previously unused) for 3 minutes each.
11. Clear slides in two changes of xylene or xylene substitute (fresh, previously unused) for 3 minutes each[6].
12. Coverslip each slide using a resin compatible with the clearant used.

---

[4] If using staining dishes and a slide carrier, the slide carrier must be of non-metallic construction.
[5] Prior to removing the slides, the container can be shaken or stirred to remove this blue band. If a shorter staining time is desired, a staining curve should be generated to guarantee consistent and optimal staining
[6] Some xylene substitutes may require longer clearing times and/or more changes.

13. The Stain will produce a dark blue color in cell nuclei. Nucleoli within the nucleus should appear as light spaces, with a dark edge, due to the nucleolar associated chromatin. Cell cytoplasm is colorless. A distinct chromatin pattern will be present in most other types of nuclei [12].

3.2.4. Hematoxylin and Eosin(H&E) Stain

A very useful property of Hematoxylin and Eosin (H&E) staining is that the images obtained are colored based on nuclear and cytoplasmic characteristics. This stain colors the protein rich collagen structures such as extracellular material with hues of pink and the nuclear and cytoplasmic regions with hues of purple and blue. Eosin Y is a normal acid (anionic dye) which, upon oxidation, results in hematin and stains by acid dyeing. Unlike the basic nuclear staining mechanisms, Al hematin might contain both cationic and anionic complexes and performs better at nuclear staining during DNA extraction from cell nuclei. The hematoxylin stains the cell nuclei distinctly from blue to bluish purple and Eosin stains other cellular elements in the tissues from pink to red. Under ideal conditions, hematoxylin should color chromatin blue and the Optical Density (depth of the color represents the visibility of small particles). Eosin colors nucleoli red. These stains are widely used in medical diagnosis due to their simplicity in staining, performance and familiarity of the color scheme. However, compared to Feulgen, H&E has not been used often in the analysis of nuclear morphometry. H&E is thus used as a gold standard, routine stain, which was initially used on tissues. In this study, prostate biopsies stained with H&E are studied in a case/control to see if similar results occur as in Feulgen stained slides. All data on Feulgen were collected beforehand, but it is predicted that H&E should lead to the same results as Feulgen analysis.

3.2.5. Protocol for H&E Staining

1. Paraffin blocks were sectioned later de-paraffinized in 2 changes of Xylene each for a span of 5 minutes.
2. Re-hydrated in 2 changes of 100%, 95% and 70% ethanol for a span of 3 minutes.
3. Sections are finished in distilled water for 3 minutes.
4. Slides are stained in Gills Hematoxylin for 8 minutes and the excess water washed off for 5 minutes and immersed in Clariffier 2 for 30 seconds.
5. The slides are again washed for 3 minutes.
6. Slides are then stained in bluing solution for a minute and washed off in running water for 5 minutes.
7. Slides dipped in 95% C_2H_5OH 10 times, counter stained in Eosin Y for 3 minutes. Slides were de-hydrated in 95% and 100% ethanol for 3 minutes and then placed in 2 changes of Xylene for 3 minutes. The tissue was cover-slipped using Permount.

3.3. Image Acquisition

Cells/tissue specimens placed on glass slides are stained using special stains to display the contrast for certain structures in them. Then the slides are then scanned are scanned by digital microscopes. Total magnification of the compound microscope is given by the product of the objective and ocular value. The resolution of the microscope is defined by the quality of the lenses, alignment of the microscope and the wavelength of light used while scanning.

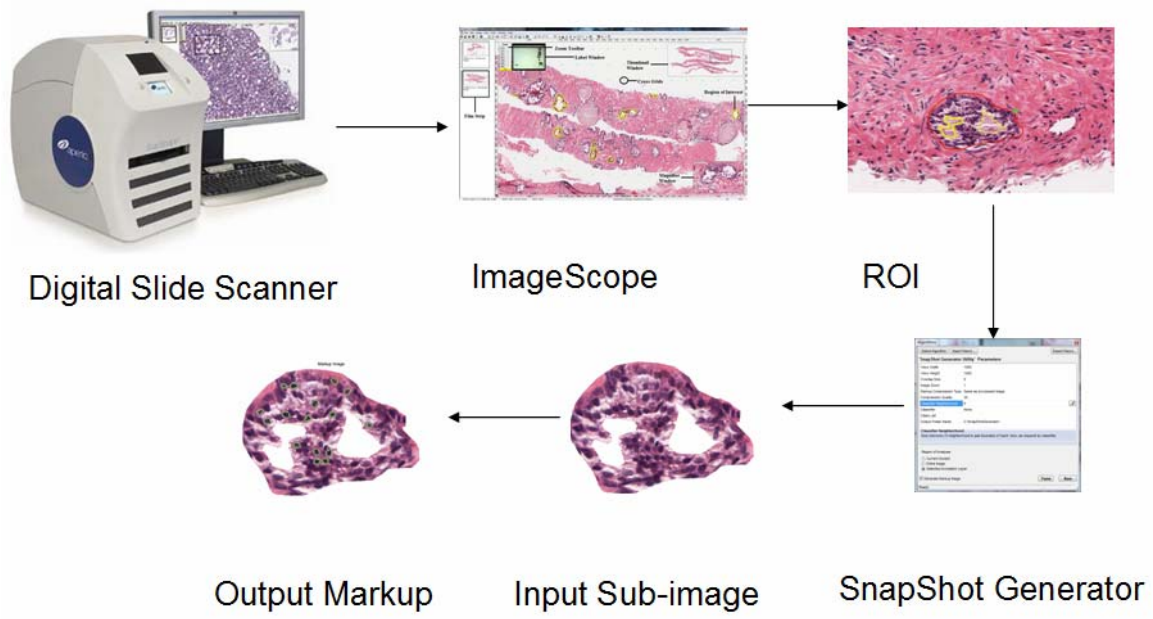

Figure 6 - Flowchart showing Image Acquisition

3.3.1. Scanning in Aperio

The stained slides are scanned using Aperio ScanScope microscope at 40X with parameter for Quality as Q80. The glass slides are scanned and saved as images on the Spectrum web server in order to be viewed from any workstation on the network, eliminating the delay in physically transporting slides. The main function of the ImageScope involves sharing and the digital slides in real time in multiple remote locations by conferencing. With the help of the various tools provided one can analyze, draw free form regions, compress, rotate and finally save the changes to a specimen image.

Annotate digital slides, making use of the following features:

a) Ability to mark the regions to be excluded from analysis.

b) Link annotations or images to create a viewing sequence

c) Add text and descriptions to annotations

d) Import and export annotations

e) Organize the annotations per user or department by creating annotation layers.

f) Instant pan and zoom capability

g) Interface the Aperio's ImageServer™ and Spectrum ™

h) Aperio algorithms support incremental processing[7]

i) Incremental processing allows the algorithm to analyze only regions added after the initial analysis without re- analyzing the previously analyzed regions

Image Scope allows working on various image formats like SVS (Scanscope Virtual Slide), JPEG, TIFF & CWS. All files under the project on Aperio Server are accessed through the Spectrum. Spectrum is a platform which allows managing digital slides under a project. The scanned images from the ScanScope instrument are stored in SVS format on Aperio Server. The images are then opened by logging into the Spectrum platform. Image Scope opens these images for further drawing and processing. The region drawing generates Annotation layers and XML files on the web server, which stores information about the layers drawn. Annotation layers are of different colors and have different names according to the type or the user. The user can also hide/show & delete a specific layer at a time. The XML files containing layers information can be exported outside the Image Scope and can be processed for further calculations.

a) The regions were drawn with free hand within Aperio ImageScope software.

b) Each of the ROIs (Oleksiy Tsybrovskyy, p.191, 1999) from the 42 patient slides was extracted as sub-images.

To acquire these regions, the XML files (containing layers information) were exported outside the spectrum. The XML files were processed using the external software

[7] Aperio ImageScope User Guide, 2011

written n C++. The software modifies the XML files to draw layers around the cancer region.

These processed XML files are stored on the server, which replaces the unprocessed XML layers; hence, next time when the image is opened, it will contain the black bounding boxes.

The compression uses a Quality factor of 80. Quality Factor is related to the compression ratio, which is defined as:

Compression Ratio = Input image size / Output image size

The quality factor is inversely proportional to the compression ratio. The regions drawn are large in size and occupy lot of memory space. Hence, these regions are then broken up into small sub-images by using extract Region tool in ImageScope. When the regions are selected, a Digital Slide Studio opens as shown in figure 7.

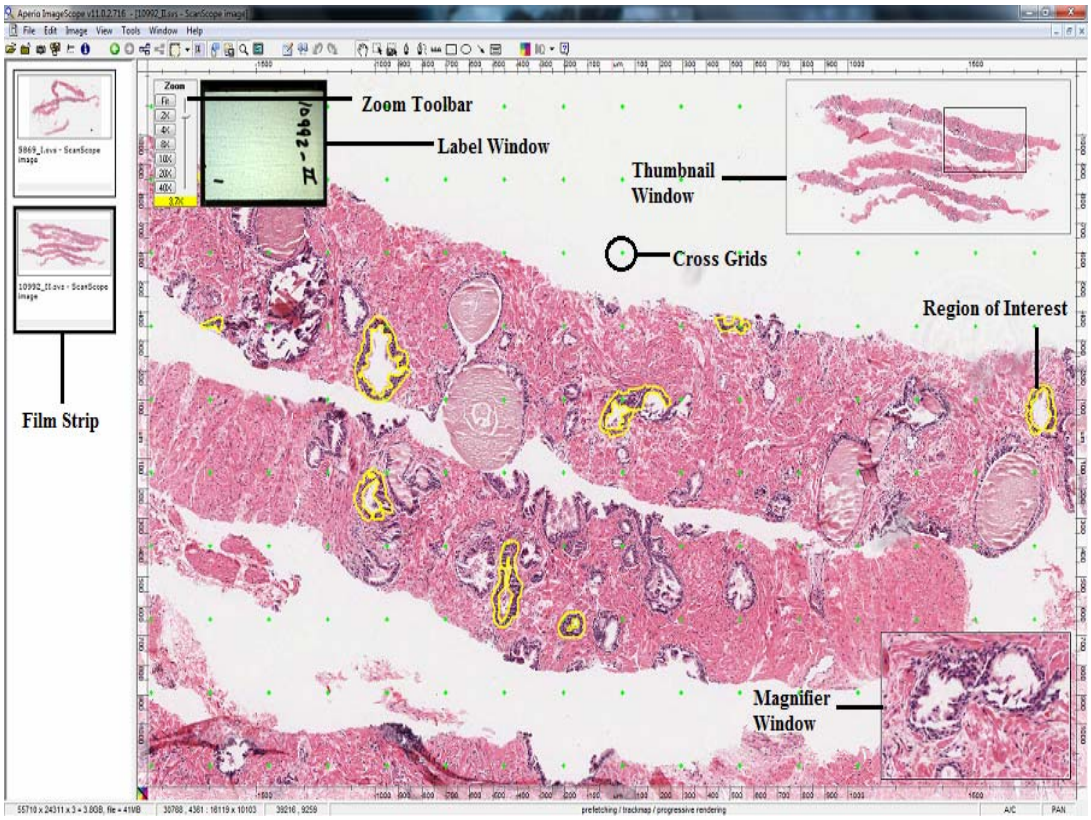

Figure 7 – Screenshot of Digital Slide Studio, an accessory which comes with Aperio software used for compression using different formats of files

The main idea behind drawing the region of interest is to make the segmentation task easier. Pathologists draw ROIs in order to exclude the lumen and stromal cells and other blank spaces on the slides. Unlike the Feulgen stained cells H&E stain makes segmentation very complicated due to the irregularity of the cell shape and texture. Also, the reason one has had to opt for generating sub-images is to save on the computation time utilized in MatLab. Even with the use of batch processing functions the processing time used is quite large. With the help of the ROI the filtering process becomes much simpler and faster.

3.3.2. Comparison between JPEG and TIFF compression

1. **JPEG**

JPEG compression is lossy compression and target quality of the image can be specified here using the quality parameter. This compression technique is very beneficial over the TIFF images as it gives us images, which occupy less memory space, and hence can be further processed easily

2. **TIFF**

The TIFF compression in digital slide studio can be done in two ways:

1. Using LZW method (Lempel Ziv-Welch) which is a lossless compression technique. This method gives a compression ratio of around 4:1, which is good enough to retain the original image information.
2. Also done without any compression by acquiring the images without compression, resulting in a raw TIFF file without any compression. TIFF compression results in good resolution image along with more pixels per nuclei. The 'pixels per nuclei' value in our study for a TIFF image is around 200 - 400 pixels per nucleus.
3. In order to avoid the compression factor the images were retained of their original size by using TIFF as the format to store the images.

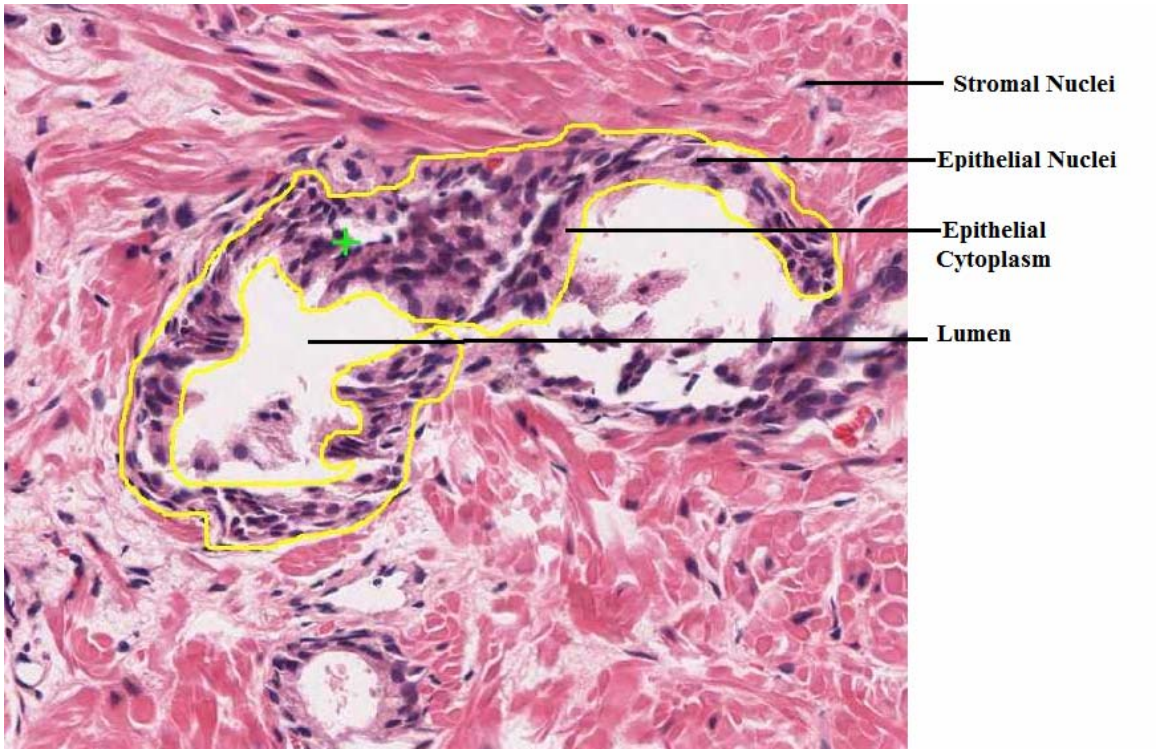

Figure 8 - ROIs (Regions of Interests) spilt into sub-images by Snapshot generator

The images obtained from the Snapshot Generator are stacked separately inside the patient folder. The size of the image varies upon the area of the region drawn. The TIFF images used for our analysis range – 150x200 to 2000x2000 pixels. Each patient slide image size > 2 GB.

3.4. Image Segmentation

The Image Acquisition step is followed by Segmentation. Segmenting epithelial nuclei in H&E-stained histopathology images is complicated by the similarity in appearance between the nuclei and other structures (e.g. stromal nuclei) in the image. Additional challenges include histological artifacts, biological variability and huge number of overlapping objects. Although active contours are widely employed in image segmentation, they are limited in their ability to segment overlapping objects and are sensitive to initialization. The Aperio ImageScope with the help of the snapshot generator

splits the whole slide image into sections/multiple images dependent on the region of interests (ROI). It generates images of TIF format loaded into MatLab for segmentation. Segmentation can be defined as a technique of separating the pixels belonging to the cell nucleus from those of the background. Manual outlining of cell nuclei is not only labor intensive but also suffers from variation due to the irregularity in the nuclear contour characteristics of the cancer cells. The chromatin level signifies the texture features, which plays a significant role in segmentation. Hence, one cannot assume an image to be a random collection of pixels for a meaningful way of interpreting them. Image segmentation falls under one of the many grouping algorithms and can be classified into region-based and contour-based approaches. While the region-based approach helps classify the image pixels into coherent image properties such as brightness, texture and color, the contour-based approach begins with an edge detection or boundary detection algorithm. A combination of edge detection applied locally to obtain the texture features makes the algorithm more realistic in biological applications.

An automated technique has lesser limitations in the number of nuclei evaluated.

3.4.1. Need for a robust segmentation technique

1. Variations in the stains within the cellular features can lead to rough margins. Though this issue is usually tackled with an appropriately sized Gaussian filter, the resultant image can no longer be used texture feature analysis.
2. Adopting a proper Cell Segmentation technique is very essential to eliminate detecting false positive nuclei. In many cases, the basal cells are falsely identified as epithelial cells.

3. Clustering of cells in glandular structures makes the process of segmentation tricky. An ideal method is required to correctly separate the overlapping objects.

3.4.2. Steps for Morphological Image Analysis

1. Reading the sub-image
2. Apply K-Means for Color Segmentation (O Sertel , p.169, 2009)
3. Apply Radial Symmetry Voting technique and detect nuclei center
4. Perform Watershed Segmentation
5. Filter objects based on morphological features

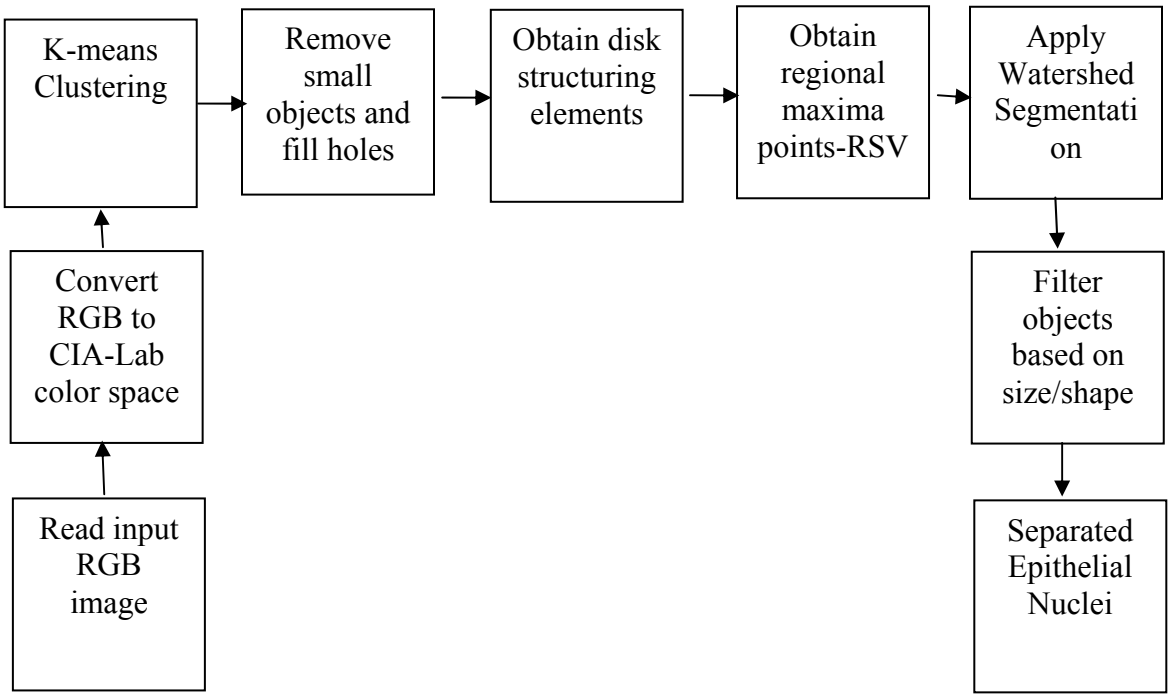

Figure 9 - Flowchart showing the steps in Image Segmentation

3.4.2.1. Reading the sub-image

H&E stained images are automatically read from the file containing all the patient folders in MatLab to segment the tissue components, glands and epithelial cell nuclei. The figure shown below is one such region drawn using the ImageScope and extracted as a single image using the Snapshot Generator.

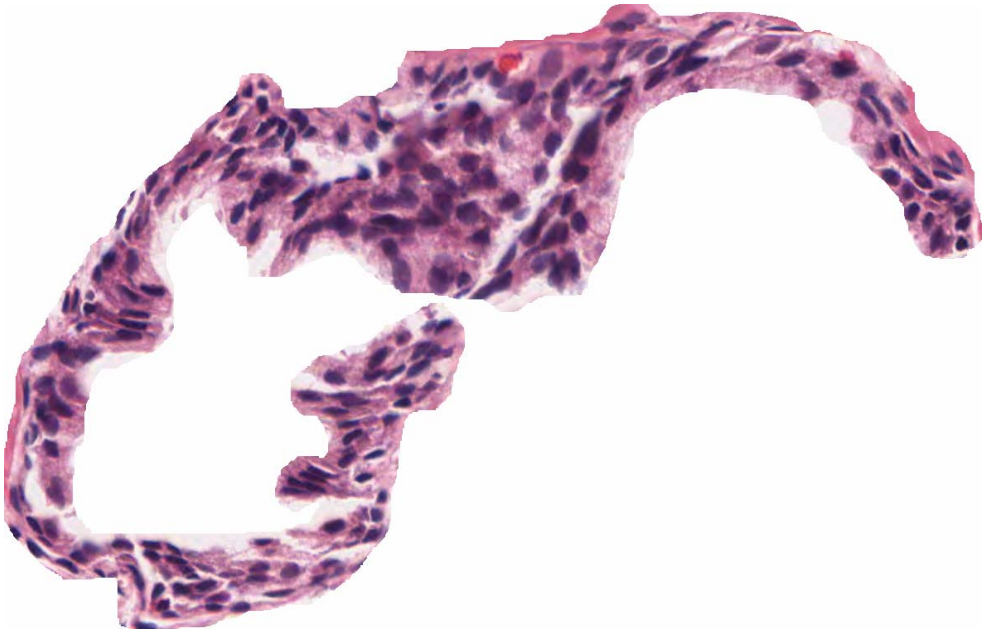

Figure 10 - Input H&E stained sub-image read into Matlab

3.4.2.2. Apply K-Means for Color Segmentation

The input H&E stained image constitutes three main colors white, purple and pink. The RGB image is converted into L*a*b* space (CIELAB), where L* indicates the Luminosity, a* the color range along red-green axis and b* represents the color ranges along the blue-yellow axis (M Recky et al, p.356, 2010). The a* and b* values also termed as opponent dimensions are measured using the Squared Euclidean distance metrics. This color space is designed to approximate the human vision with its

perceptual[8] uniformity. The component L matches the lightness perception by humans, thus allowing it to be used for accurate color balance corrections. Since our main analysis includes texture feature extraction it is essential to retain the color information instead of converting directly to gray scale format for segmentation. Also the K-means algorithm used in the next step produces better outcome with the color information. K-Means is a classic unsupervised learning method for partitioning the data into k mutually exclusive data sets depending on the color values generated by *a* and *b*. Hence, for ideal classification it is important that every pixel in the image has a value in space. The *a* range segments the Hematoxylin stained nuclei while *b* value segments the white regions, which include both the background and the lumen. The clustering algorithm works on the principle of maximizing the between class scatter (S_b) matrix and minimizing the within class scatter matrix (S_w). The Squared Euclidean measure implemented here utilizes the centroid to be the mean of the points in that cluster is found to work best on 2-D vectors in comparison with City Block, Cosine, Correlation and Hamming measuring metrics. Due to the unusual intensity of the stain used in the data sample the algorithm was set with k = 2. By using Lab space the color information was well utilized for segmentation. The K-Means Clustering algorithm gives the background markers used as an input for Watershed segmentation.

[8] Perceptual means the same amount of change in color values produce the same amount of perceptual difference of visual importance

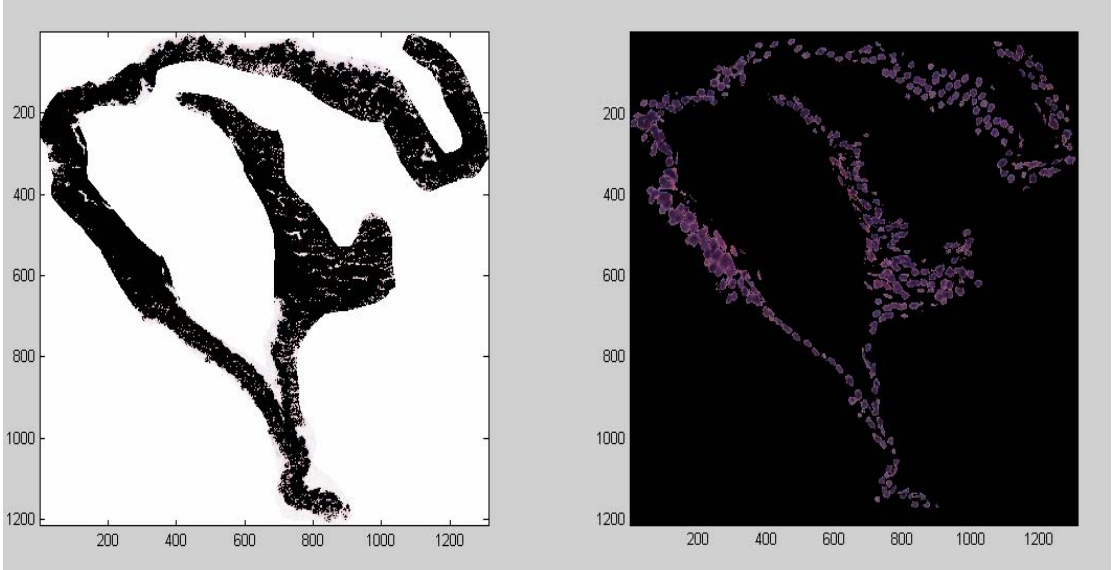

Figure 11 - Foreground and background markers by K-Means

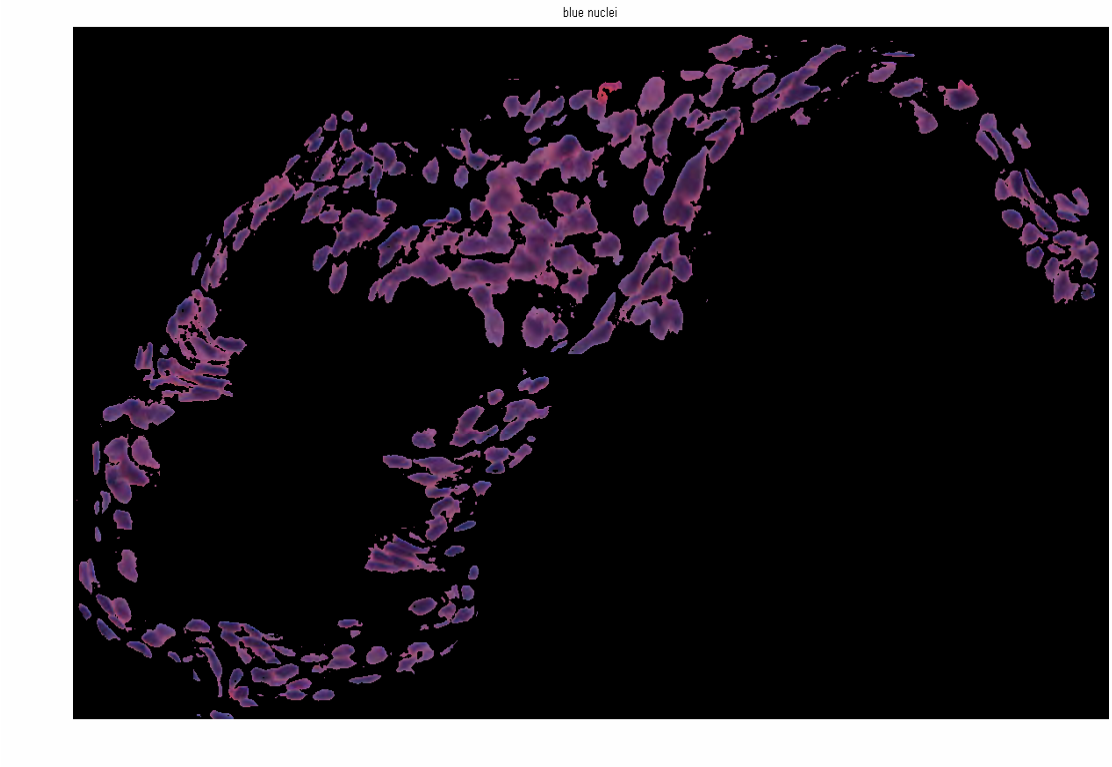

Figure 12 - Image showing the result of color segmentation

3.4.2.3. Apply Radial Symmetry voting technique and detect the nuclei center

In this technique, the centre of mass is given the maximum value and called the local maxima. Radial or tangential symmetry (O Schmitt, p.1905, 2005) works along the boundary and is finally reduced to a single isolated point or at least a closely associated group of points. With continued refinement, a single focal output as the nuclei center is obtained. In many substructures the nuclei contains overlapping boundaries due to the presence of chromatin. With no prior information about the object (nuclei) locations, it is safe to assume that the centre of mass concentration is along the radial direction. By the voting based technique, we mean the number of iterations carried out to reduce the angular range from a large value to almost zero until the nuclei center is detected.

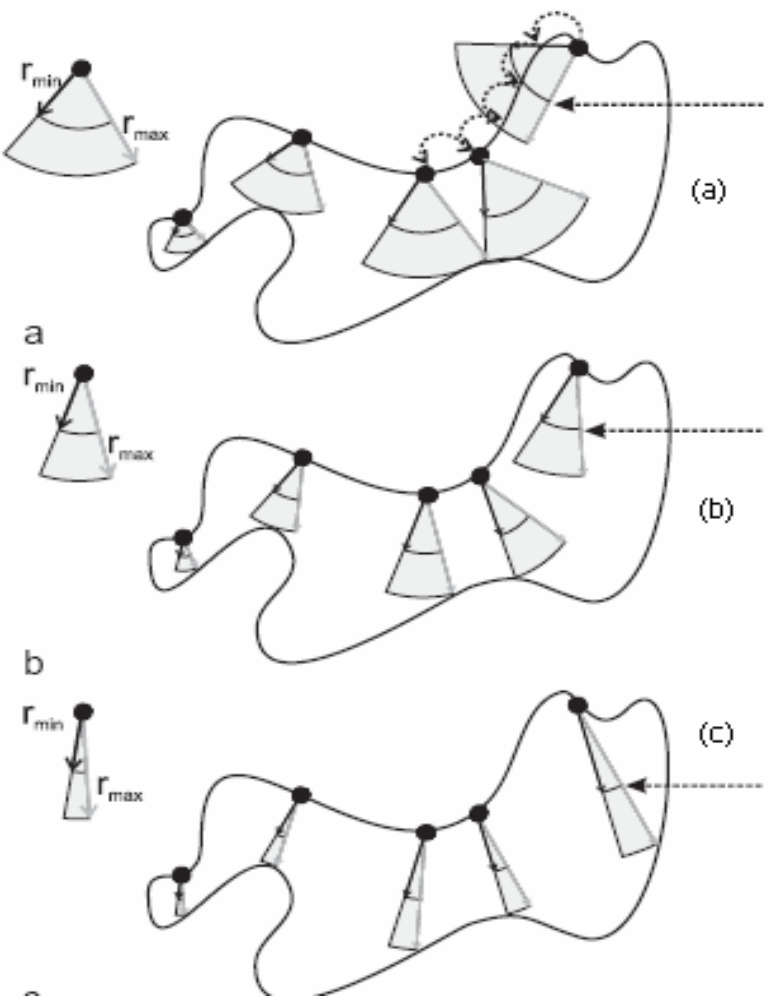

Figure 13 - Orientation of kernel obtained at each step of the Radial Symmetry based Voting technique in order to find the nuclei center

P is each point in the image

Q is the maximum in P's voting area

(rmin, rmax) - radial range

Δ - angular range

Determining the step size is essential in determining the voting area. Larger the step size means the center mass undergoes more fragmentation. The radial symmetry based voting technique acts as a useful input to the watershed segmentation

procedure and it can be applied to both dark and bright regions. One can perform this technique on image gradients or binary images displaying the gravity center. If the chromatin material is evenly distributed with respect to the nuclei center, then the center mass is zero. Since the nuclei are radially symmetric objects this operation is suitable for their localization. The Sobel gradient magnitude is used as a segmentation function to obtain the structuring element such as a disk on which radial symmetry is performed. The foreground marker behaves as the regional minima.

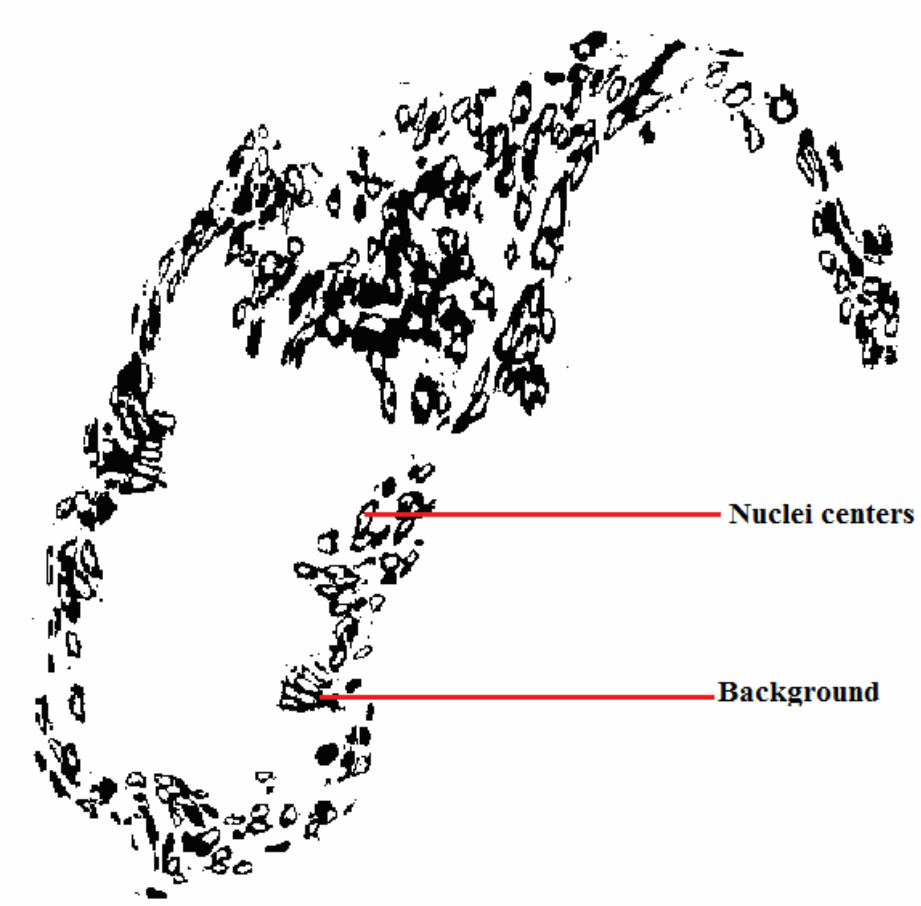

Figure 14 - Binary Image of Nuclei Centers from (RSV) Radial Symmetry Voting-based technique

The white spaces inside the nuclei correspond to the detected centers and the remaining portion of the nuclei is black.

3.4.2.4. Perform Watershed Segmentation

Segmenting an image using watershed involves morphological information. The watershed transformation (Malik Khan, p.546, 2009) largely relies on the gradient magnitude of an image to categorize it into topographic surfaces. Low-contrast edges produce small magnitude gradients, causing distinct regions to be erroneously merged and resulting in under segmentation. Over-segmentation occurs when the watershed results are degraded by background noise. Watershed segmentation is a region-based technique applied to gray scale images using flooding process.

The classical watershed segmentation technique can be modified by varying the method used to obtain the markers. Watershed can be explained in terms of the valleys peaks and basins. The basins form the regional minima while peaks form the maximal points. The minima points are the nuclei centers used as foreground markers. In case of H&E stained images the objects tend to be overlapped causing trouble during segmentation of the nuclei. Leading to inaccurate detection where in many objects might belong to single regional minima.

Though it separates the overlapping and touching nuclei, the procedure does so by sometimes producing unsatisfying splitting (parallel to the borders). The results generated will not be accurate enough to map the entire shape of the nuclei. The inputs to the WS technique are EDM obtained via the K-means algorithm.

3.4.2.5. Filter objects based on morphological features

Upon segmentation, the image is further filtered based on the shape and size features (Area, Perimeter and Circularity). With the help of a lower and upper threshold value set to each features the desired epithelial nuclei were obtained. Currently, the algorithm has a limitation on the number of nuclei detected since it works closely on the removal of false positives. The figure shows the markup image of the final segmented nuclei marked by the green boundary.

Figure 15 - Markup image of the segmented epithelial nuclei

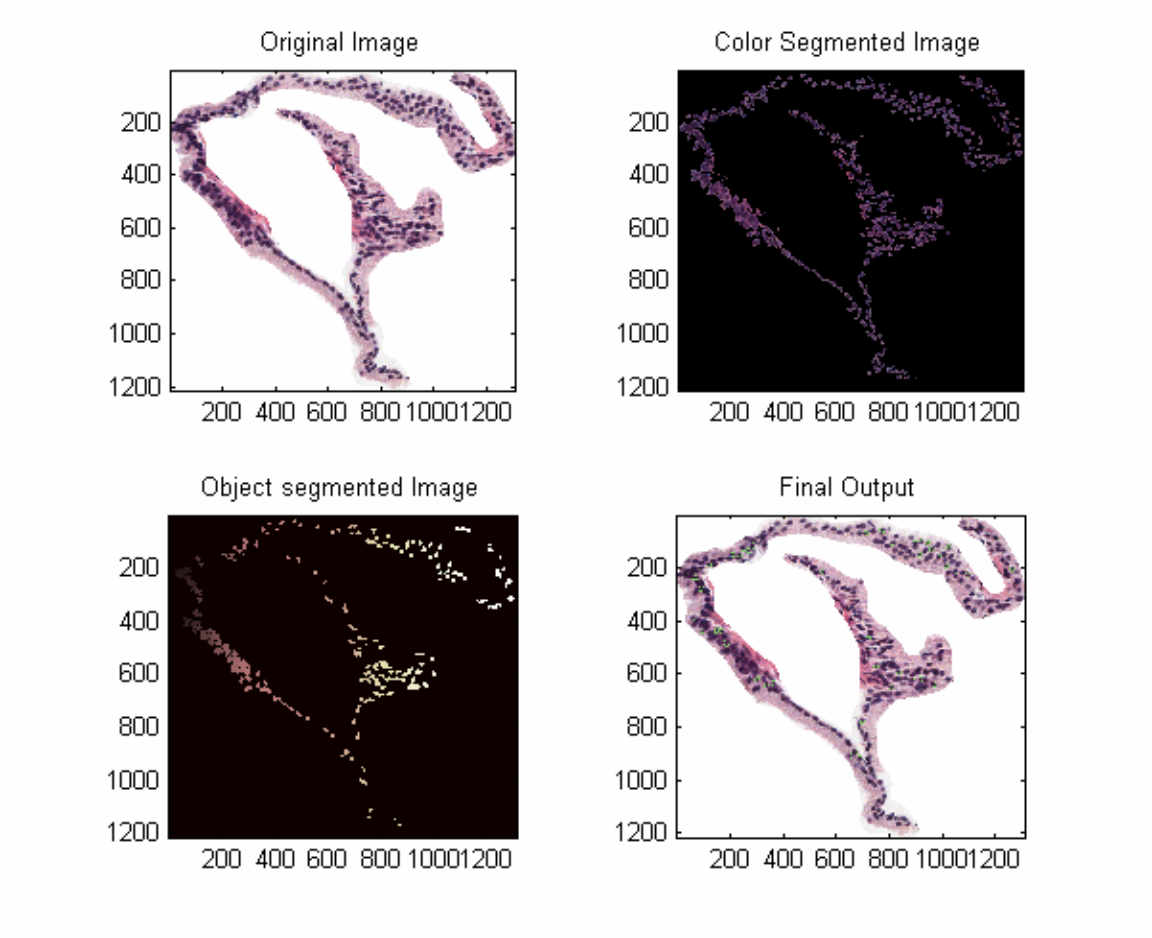

Figure 16 - Segmentation steps for a sub-image

3.4.3. Batch to Batch Staining Variability

Although our study included no variations in the stains, in order to avoid the batch to batch variation in stains, the data was batch normalized. Z-scores are generated for each nuclear image in the high-grade cancer galleries by subtracting the mean of the normal nuclei for that feature from the raw measure for the nucleus in question and dividing by the standard deviation for the normal nuclei. The values with negative z-distributions (shifted to left of the normal), z-score distributions are close to zero mean and distributions with low variance.

3.5. Feature Extraction

The main purpose of selecting the epithelial nuclei is to be able to find the morphological features which can prove useful in the PCa prediction. An extensive set of features are constructed from this intermediate representation in order to characterize the tissue. We observe the architectural and texture characteristics of the tissue under study. A set of 180 features from the shape/size, DNA content, Markovian, texture feature database was chosen for the current study. Each segmented nuclei was characterized based on all the features in MatLab and the output was stored as a comma separated (csv.) file format.

The resultant output file was further read into SAS for analysis.

Table IV - Upper and lower threshold levels used for feature extraction

Value	Upper threshold	Lower threshold
1	0	1
2	0	1.5
3	0	2
4	0	2.5
5	1	1
6	1	1.5
7	1	2
8	1	2.5

4. STATITISTICAL DATA ANALYSIS AND MODELING

The mean values of the cell features over the individual slide as well as the standard deviation of feature measurements within the slide are calculated for all the individuals. The slide based population distribution was examined and compared for each feature over the entire group of patients. With the help of backwards selection technique only 8 features make into the final model.

Table V - Final Features Selected for Analysis

Shape/Size	Texture	
Perimeter Circularity	Sum of OD Average of OD FeretY Slope 4 [9] TMean TSD TSmooth TEntropy Max OD Sum Mean Sum Variance Sum Entropy Diff Energy Cluster Shade Cluster Prominence Contrast Homogeneity	2nd Angular Moment Entropy Low DNA Area 2 [10] Medium DNA Area2 Low DNA Amount2 Low DNA Compactness2 Medium DNA Compactness2 High DNA Compactness2 Low Center Mass2

[9] Lower threshold = 0,Upper threshold = 2.5
[10] Lower threshold = 0,Upper threshold = 1.5

4.1. 2-Step Model used for the statistical Analysis: Logistic Regression

Linear logistic regression or logistic model (So Young Sohn, 2007, p.472) uses various predictor variables to derive a scoring function that can distinguish the cases versus the controls. Feature measurements for around 200 cells from each patient were pooled to form the groups. Each feature was then normalized to the mean and standard deviation of the total number of nuclei. Leave-one-out (LOO) cross-validation was performed on the same dataset to verify the fit of the model. Each time with holding one out of the 42 patient data set. Out of the total 180 features, which also included the threshold features, only 28 of them were made into the model after applying backwards elimination to the selected features. The resulting observations were used to calculate the Multi-Feature Score on a nuclear level (nMFS) derived by weighing 28 features using the regression co-efficient. The nMFS plotted as an exponential curve gives a nuclei probability of becoming malignant in future. Indices such as mean and standard deviation were used to characterize the shape of the frequency distribution of nMFS within each patient image. Once again, linear logistic regression analysis generates the summary statistics of nMFS using the regression co-efficient to derive the person level Multi-Feature Score (pMFS). The area under the curve was used as a measure of discriminating the power of the pMFS model. The box plot plotted as a comparison between the case and control for the H&E dataset.

5. RESULTS

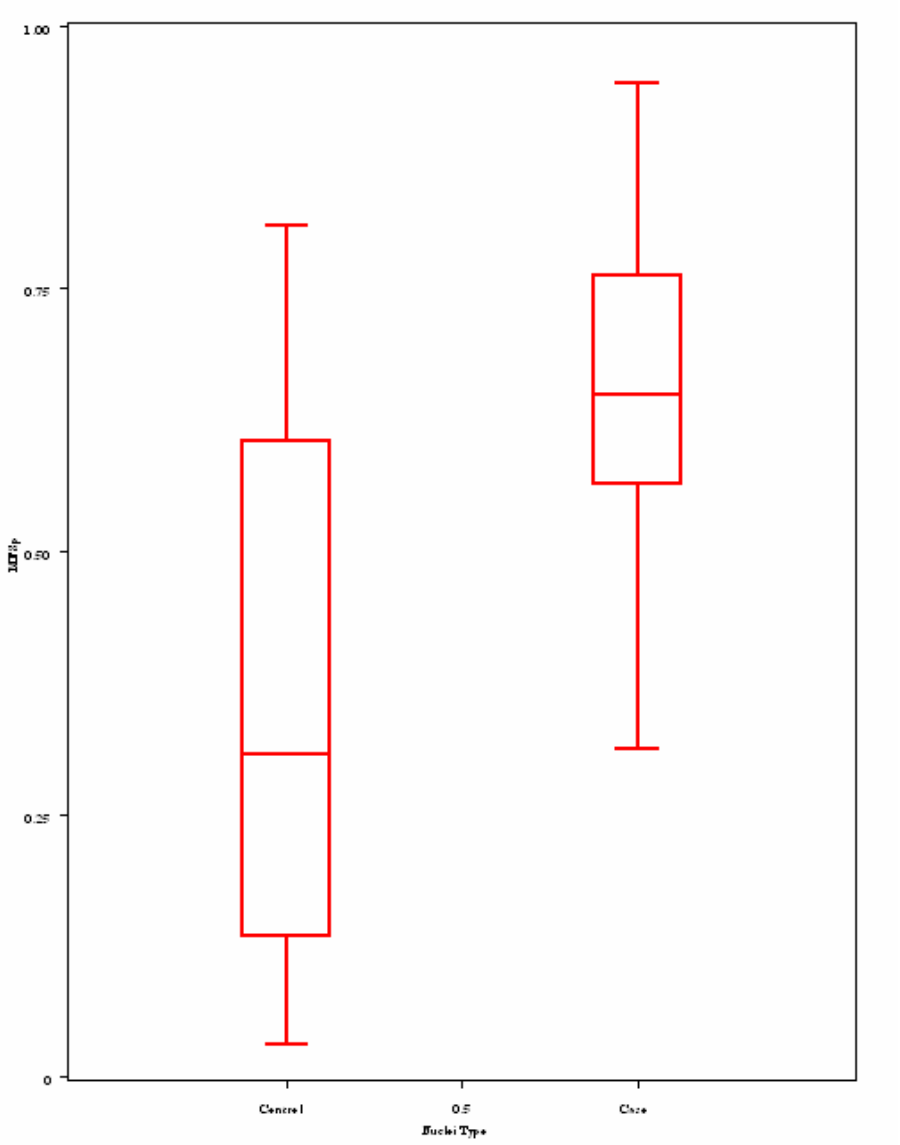

Figure 17 - Box plot of pMSF H&E (Cases and Controls)

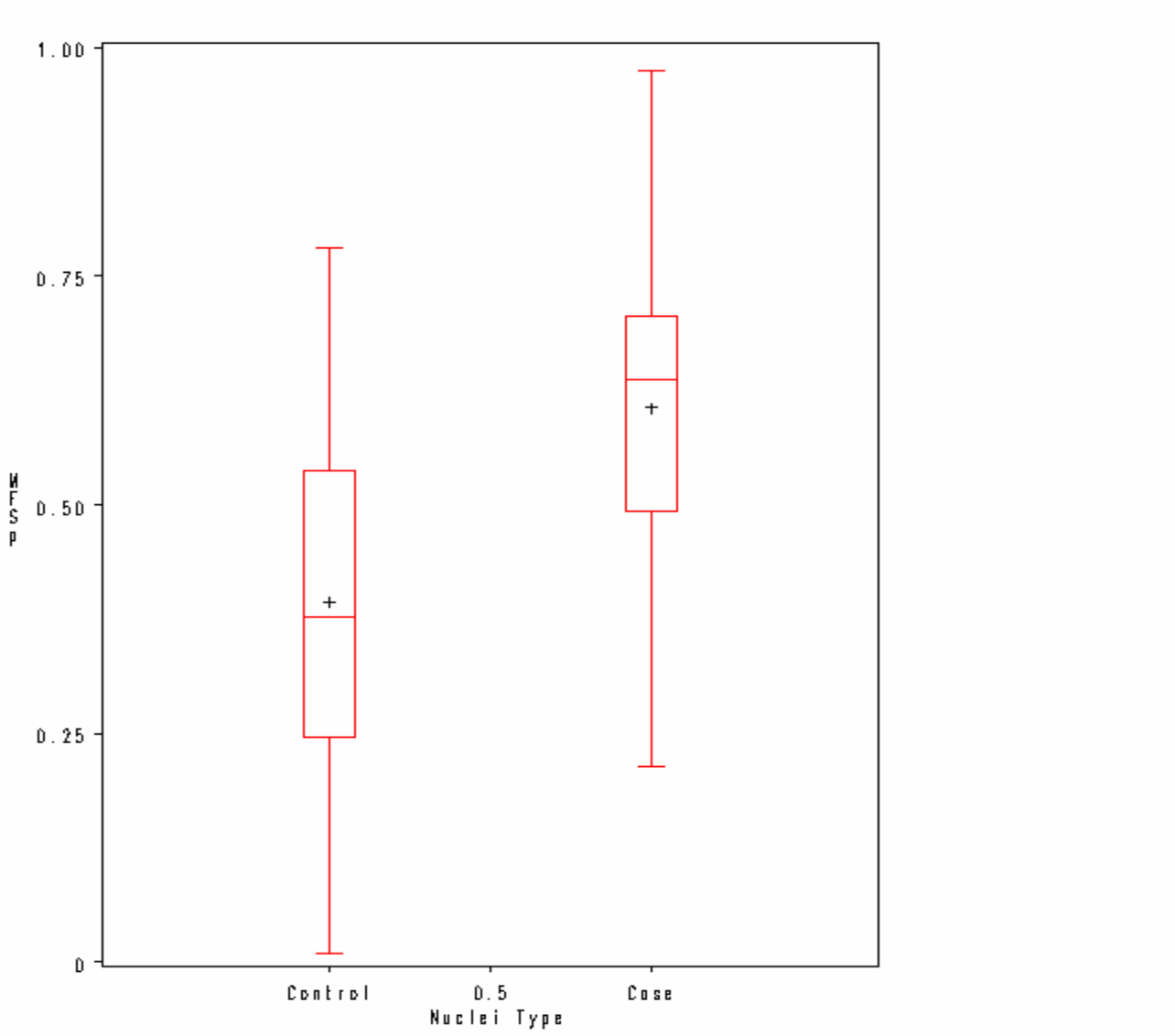

Figure 18 - Box plot of pMSF Feulgen (Cases and Controls)

5.1. ROC

The Receiver Operating Curves (ROCs) provides non-parametric comparison of areas under uncorrelated ROC curves. It provides point and confidence interval estimates of each curve's area and of the pair-wise differences among the areas. Tests of the pair-wise differences are also generated. Any contrast among the areas may be estimated and tested.

Using the predicted probabilities of the LR model for each case, ROC analysis was performed to identify the ability of the models to predict the development of cancer using the morphometric mean and StdOD. ROC is plotted with specificity on the x-axis and sensitivity on the y-axis. Studies based on morphometry should be interpreted with caution, as the data is inherently redundant and multivariate in nature owing to the large number of correlated predictors. The Area under the receiver operating curve (AUC) was computed to be 0.77 showing that our analysis of the cases and controls was 77% appropriate.

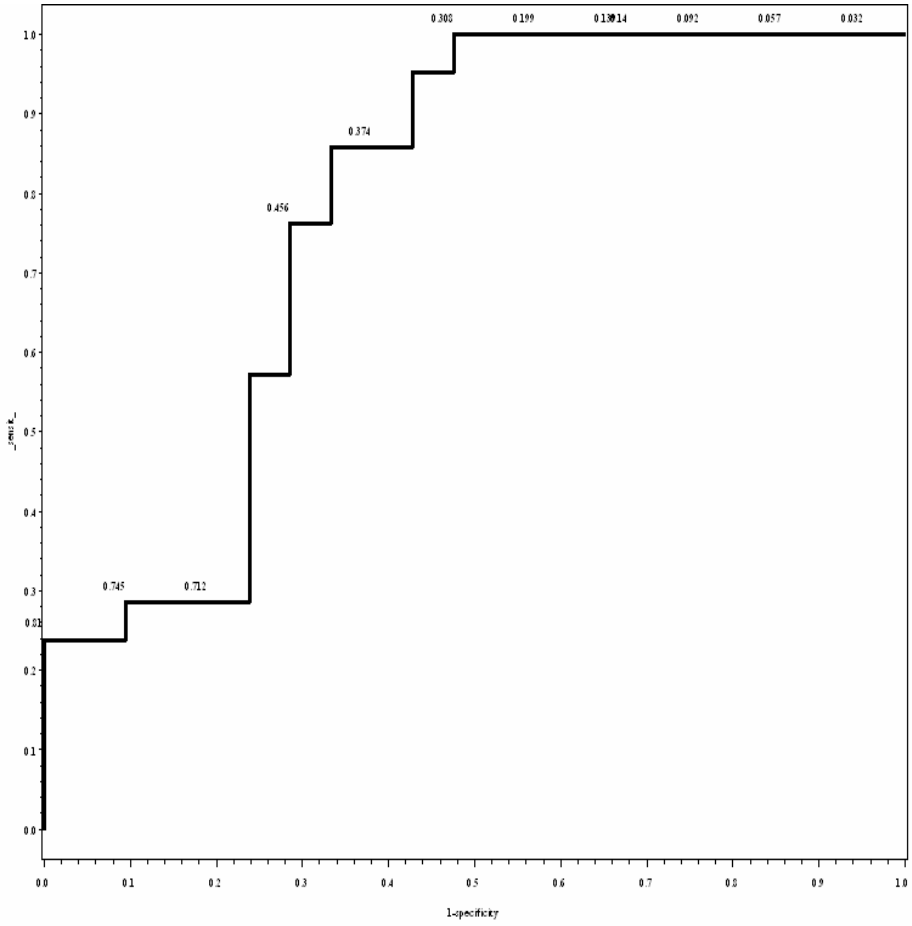

Figure 19 - Receiver Operating Characteristics from the 2 Step Model

5.2. Paired t-tests

The accuracy of the automated segmentation was also verified by the paired t-test conducted over all the 42 datasets for the selected features. P values have been considered as parametric estimates based on normal distribution, $P < 0.05$ is taken as significant. The number of samples was found to be less to generate a valid diagnosis leading to over-fitting. Upon conducting the paired t-test P-value obtained for H&E person level MFS was 0.000234.

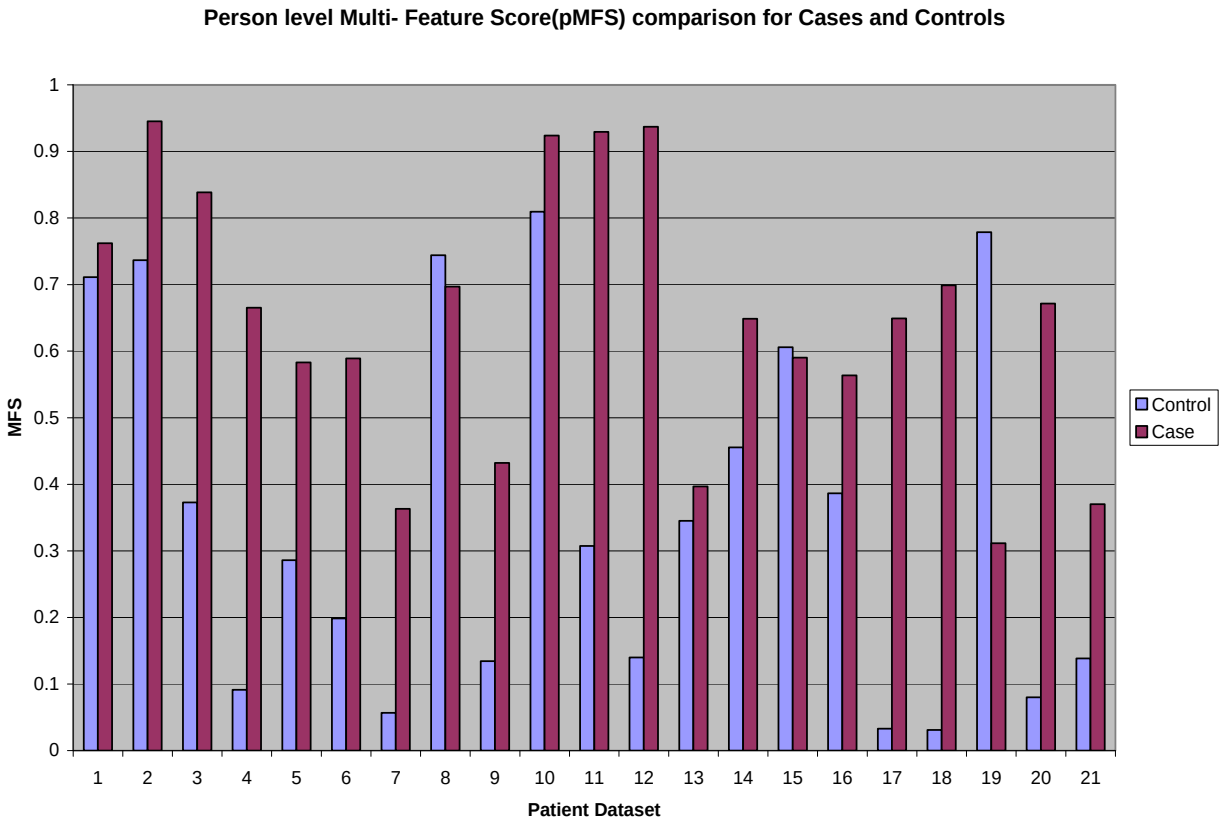

Figure 20 - pMFS comparison for Cases and Controls H&E

Another form of verification is the Pearson or Spearman correlation. The Spearman is a nonparametric correlation based on ranking the two variables while Pearson is based on the assumption that variables are obtained from sampled Gaussian distribution (i.e., ties in the original values). Spearman rank correlation is also used when the data do not meet the assumptions about normality and linearity.

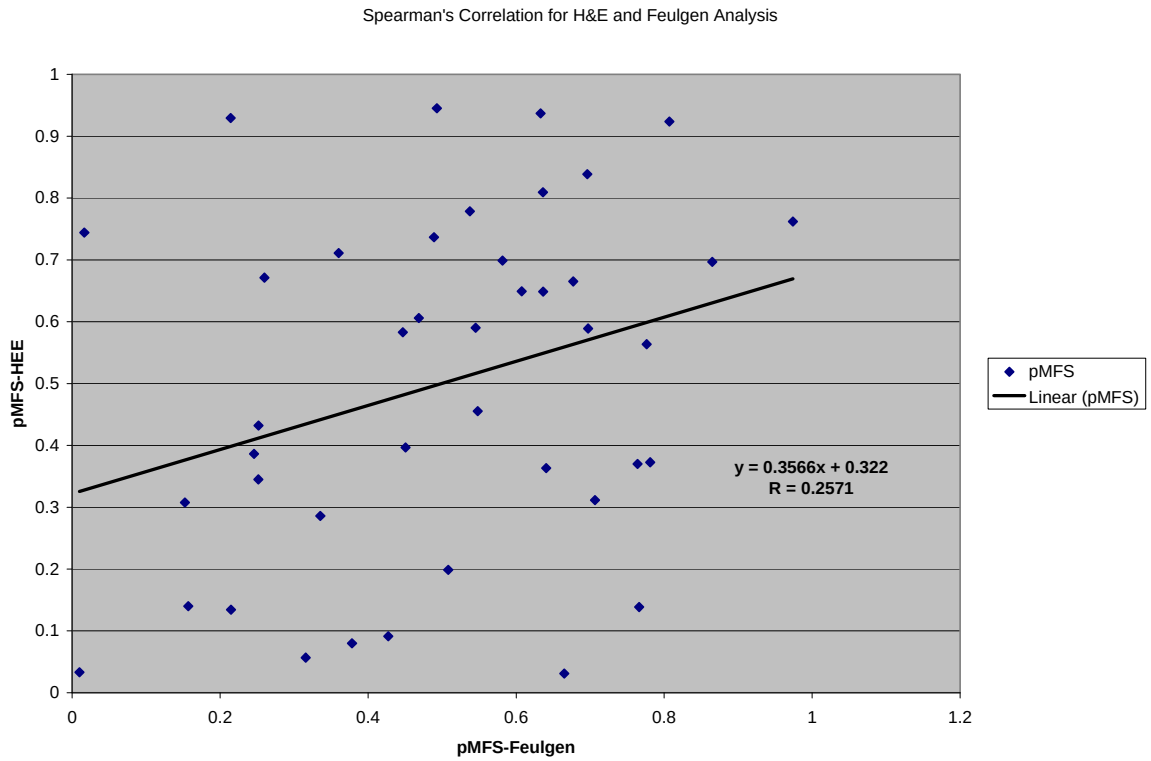

Figure 21 - Scatter Plot for Spearman's Correlation Analysis for H&E and Feulgen pMFS values

While for the case versus controls in the H&E data we get R = 0.1779, N = 21 but the P-value obtained here is insignificant since the number of samples considered is negligible.

5.3. Selective Bi - variate Comparisons for Feulgen versus H&E sample set

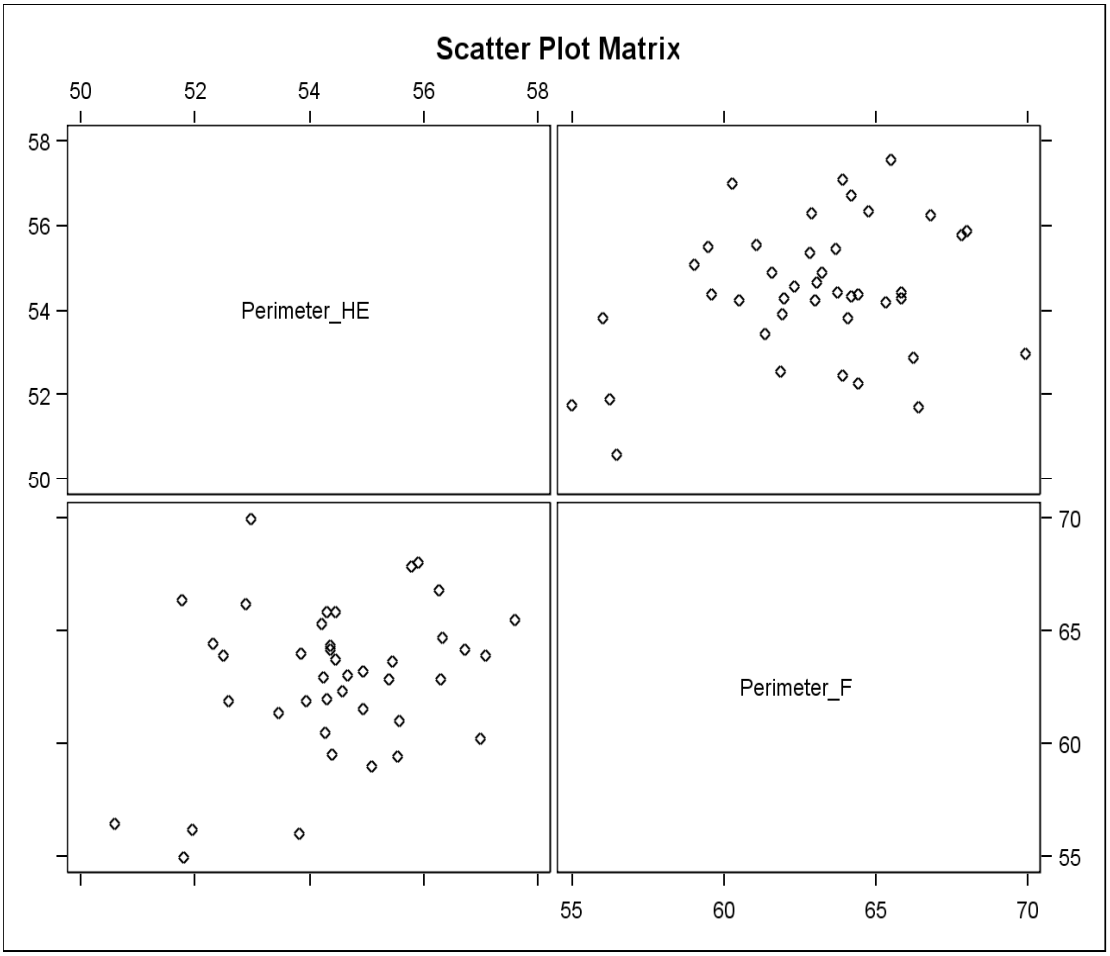

Figure 22 – Size (Perimeter) Feature wise Correlation for H&E and Feulgen Stain

Pearson Correlation Coefficients, N = 41 Prob > |r| under H0: Rho=0		
	Perimeter_HE	**Perimeter_F**
Perimeter_HE	1.00000	0.29759 0.0588
Perimeter_F	0.29759 0.0588	1.00000

Spearman Correlation Coefficients, N = 41 Prob > |r| under H0: Rho=0		
	Perimeter_HE	**Perimeter_F**
Perimeter_HE	1.00000	0.16707 0.2965
Perimeter_F	0.16707 0.2965	1.00000

Figure 23 - Spearman and Pearson Correlation specific to Size Feature (Perimeter) for H&E

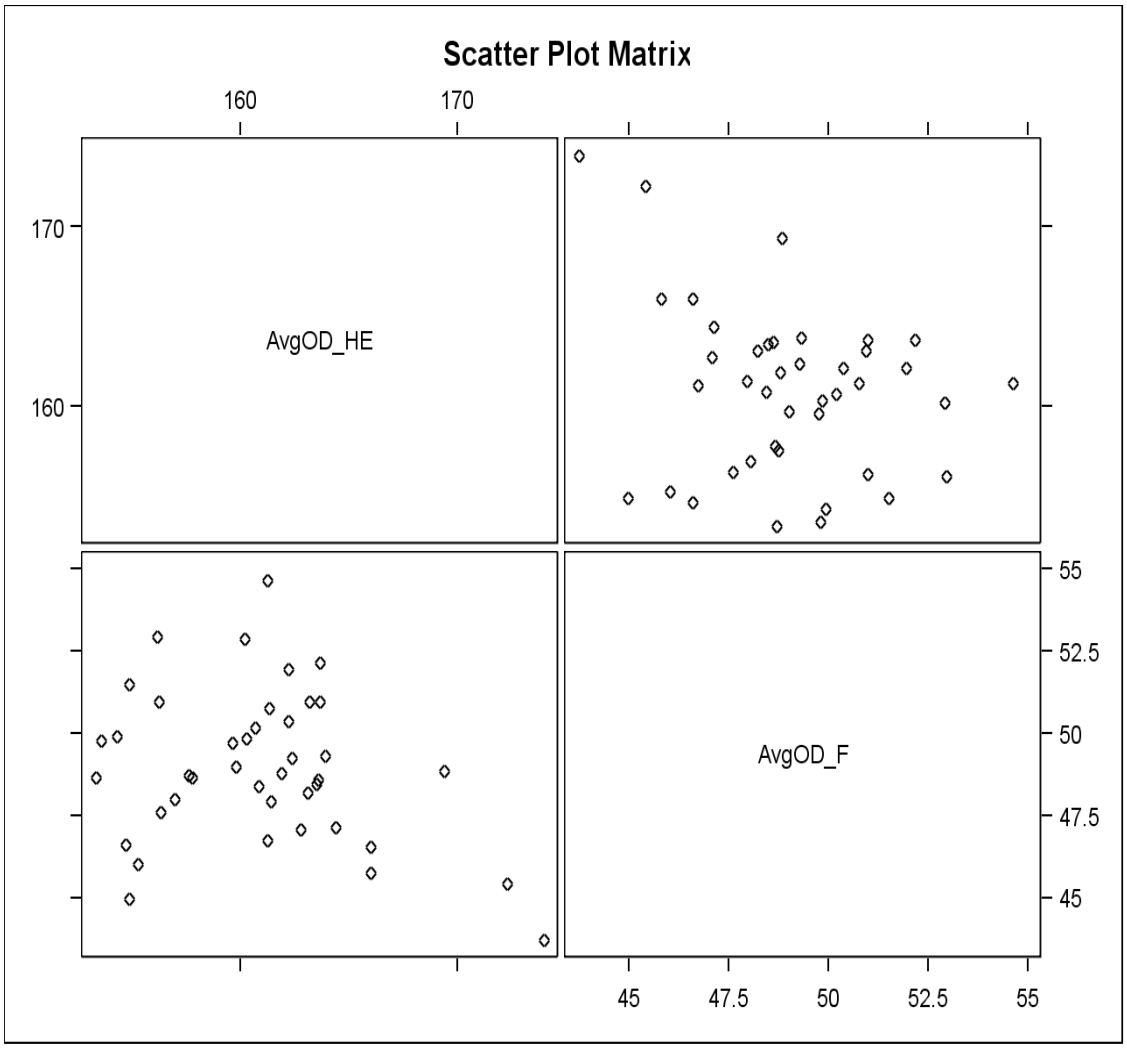

Figure 24 - Texture Feature-wise Correlation for H&E and Feulgen Stain

Pearson Correlation Coefficients, N = 41 Prob > |r| under H0: Rho=0		
	AvgOD_HE	AvgOD_F
AvgOD_HE	1.00000	-0.26555 0.0933
AvgOD_F	-0.26555 0.0933	1.00000

Spearman Correlation Coefficients, N = 41 Prob > |r| under H0: Rho=0		
	AvgOD_HE	AvgOD_F
AvgOD_HE	1.00000	-0.17247 0.2809
AvgOD_F	-0.17247 0.2809	1.00000

Figure 25 - Spearman and Pearson Correlation specific to Size Feature (Perimeter) for Feulgen

There seems to be no correlation between the stains (H&E and Feulgen) as seen in case of Perimeter and Average Optical Density feature. One of the main reasons for this distortion could be the difference in staining techniques. Another possible reason could be the difference in the nuclei selected.

5.4. Odds Ratio Estimate for the final H&E features

Below is a forest plot of the odds ratio measured for the final feature data set obtained from SAS. The ends of each feature represents the upper and lower limits(95% Confidence Limits).The odds graph is split into 2 subsections for better understanding.

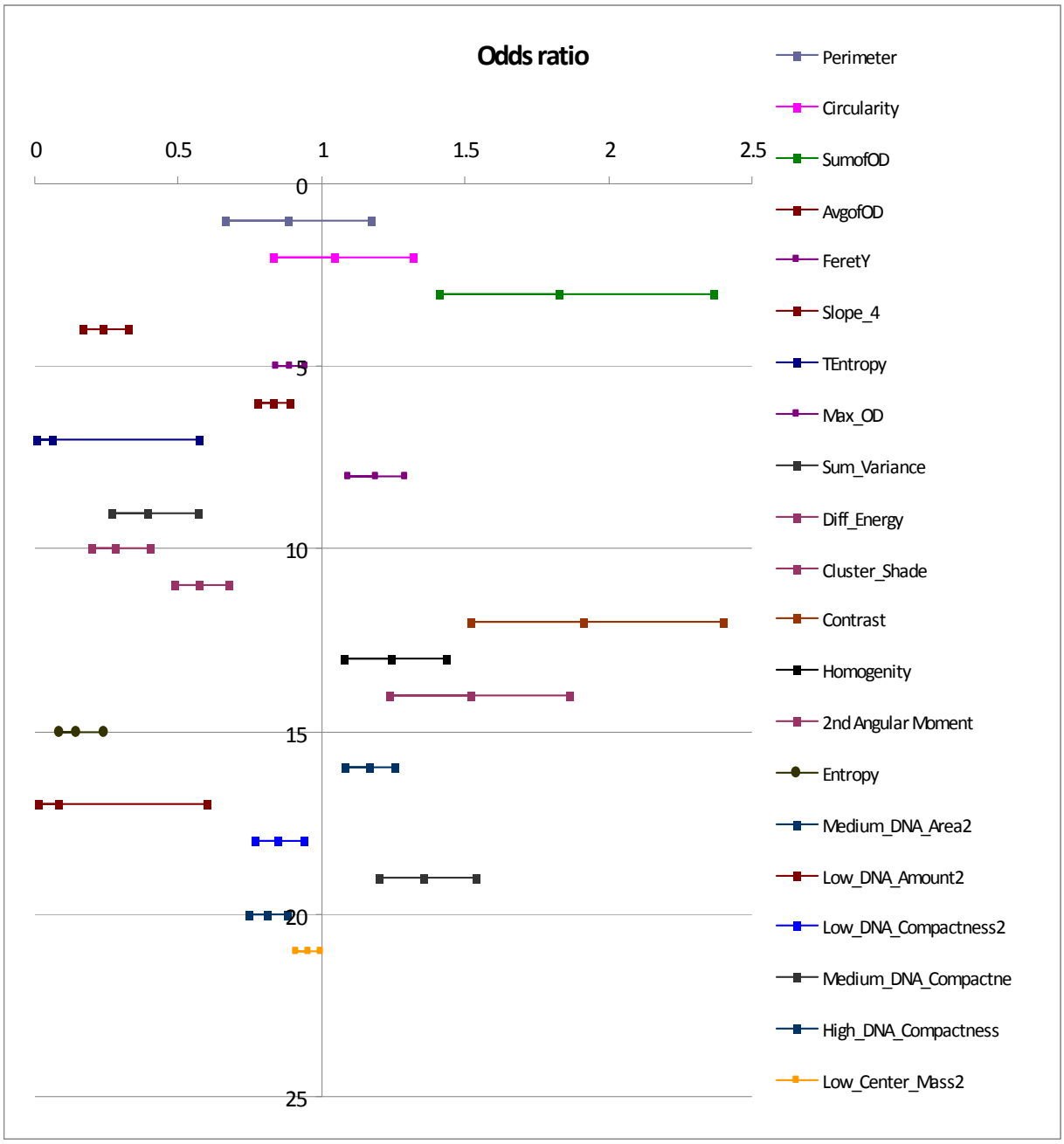

Figure 26 - Odds ratio as a forest plot for the final selected feature set 1

Features such as Perimeter, Circularity can be eliminated based on the odds ratio analysis. Importance is given to the features with lesser range between the confidence limits such as Slope4, Average OD and Low Center Mass.

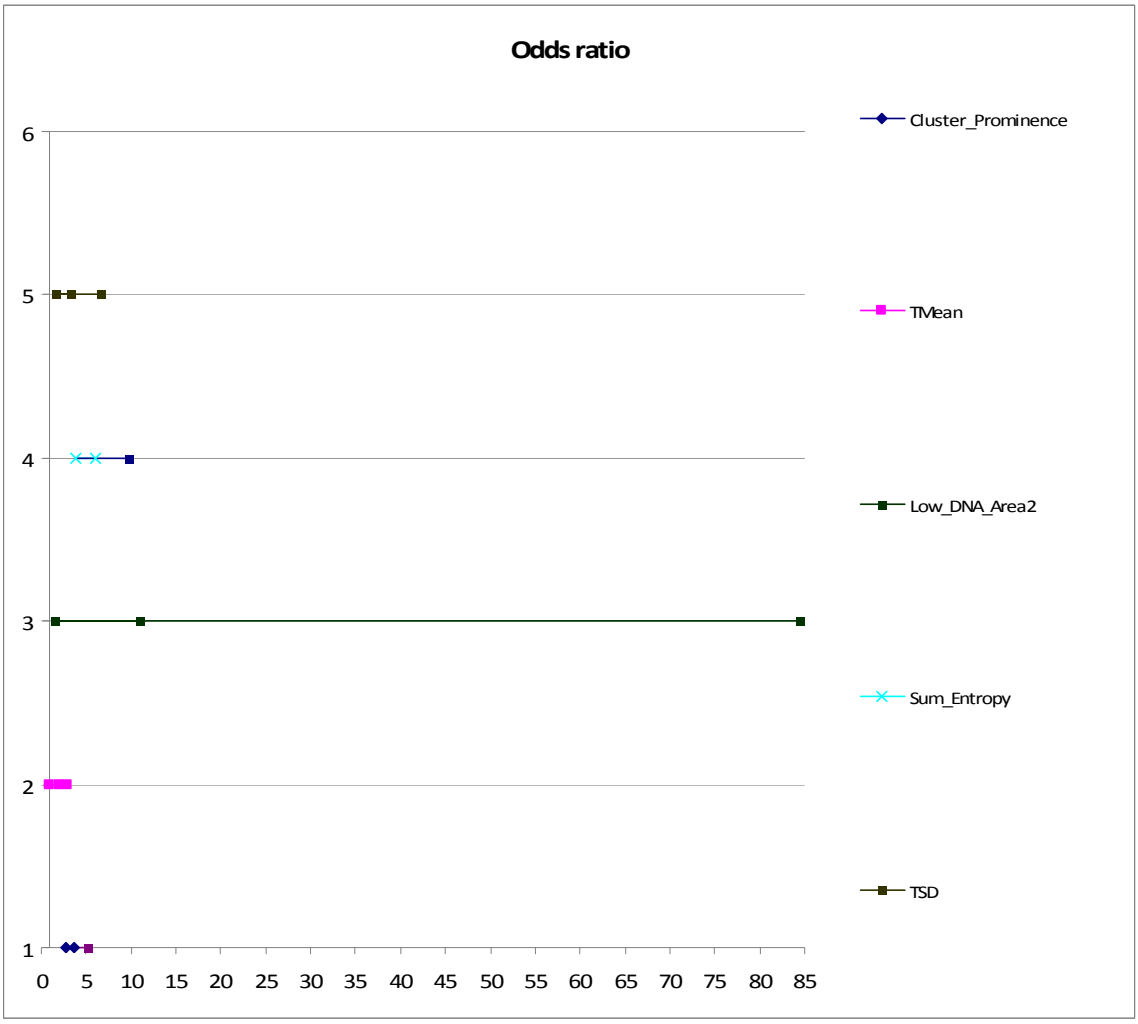

Figure 27 - Odds ratio as a forest plot for the final selected feature set 2

6. DISCUSSION

6.1. Summary and Contributions

The thesis addresses the problem of analyzing prostate cancer biopsy image data. It describes an unsupervised semi-automated method for segmentation of nuclei in Hematoxylin and Eosin (H&E) stained Prostate Biopsy images and investigates its use in modeling the data and its effectiveness predicting cell alterations. Existing methods have largely focused on the use of Feulgen-stained prostate cancer biopsies in the analysis of Nuclear Morphometry due to its DNA staining capability. In this thesis the potential for the use of the more easily available H&E staining is investigated since H&E stains are widely used in medical diagnosis due to the simplicity in the staining procedure. Our results provide evidence that the H&E can yield a performance comparable to a reference method that uses Feulgen-stained biopsies.

Earlier work on nuclear morphometry strategies had largely focused on assessing just one feature which included only OD (optical density). However, after major advances in this area many groups have tried to study various sets of nuclear morphometric features for different types of cancers, like skin cancer, esophagus cancer, bladder cancer, and prostate cancer. This algorithm used optical cellular selection for rapidly scanning a slide and identifies 180 “field of views” of which few are reviewed by pathologists before analysis. It is seen that nuclear morphometry when used with other criteria like PSA, age, and dieting habits can yield better prognostic value in stratifying the risk of the patient.

Nuclear morphometry has also been used in cervical and breast cancers to measure nuclear grade in chemoprevention setting. These procedures use histopathological grading by providing objective and quantitative assessments to the pathologists.

The segmentation technique used is well-suited for H&E stains since it successfully reduces the number of false positives in the cell image. Many a times, the nuclei and the surrounding structures are of the same color, making segmentation much trickier. With the use of radial symmetry based voting technique, one is able to clearly identify the nuclei centers threshold based on intensity. Accuracy of the segmentation of glands and cell nucleoli was assessed using manual outlines. The automated algorithm created was able to generate about 150 nuclei each for each patient. The textural and architectural features were calculated and used for statistical analysis. Area under the curve (AUC) value was equal to 0.77 on independent data sets N = 42 with promising features were combined by LR classifiers using the LOO cross validation method, on the 2nd step model to classify the difference between the case and controls. This dissertation provides preliminary evidence towards classifying the cases and controls accurately in digital histological tissue sections with H&E stains.

6.2. Recommendations for Future Work

One of the many limitations is analysis on the data set is a direct technique not containing any external samples might have affected the results obtained. For future application of the segmentation model one might want to test on samples other than cases

and controls. Due to the split sample validation on the data, a reduced set case was generated leading to an almost over-fitting model with unacceptable power. The issue can be addressed in future by using more subjects for the diagnosis. Secondly, the drastic reduction in the number of nuclei segmented has been satiated by the elimination in the false positives generated in the earlier versions of the algorithm. The next step in research would be to implement nuclear segmentation to a more diverse data set and also apply to other forms of cancer. With further development in image analysis, classification and feature extraction techniques, better prediction of the development of prostate cancer can be achieved. We believe that field effects have not been studied before extensively using nuclear morphometry and had a lot more potential in epidemiological research.

CITED LITERATURE

Bacus, J. W. and Grace, L. J., "Optical microscope system for standardized cell measurements and analyses," *Applied Optics*, Vol. 26, No. 16(1987).

Doudkine, A., MacAulay C., Poulin N., Palcic B., "Nuclear texture measurements in image cytometry". *Pathologica*, 87(3): p. 286-99, 1995.

Fatakdawala, H., et al., "Expectation-maximization driven geodesic active contour with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1676–1689, 2010.

Guillaud, M., et al., "Nuclear Morphometry as a Biomarker for Bronchial Intraepithelial Neoplasia Correlation with Genetic Damage and Cancer Development," *Cytometry Part A* 63A:pp. 34-40, 2005.

Hafiane, A., Bunyak, F and Palaniappan, K., "Fuzzy Clustering and Active Contours for Histopathology Image Segmentation and Nuclei Detection", in *Proc. ACIVS,* 5259:903-914, Oct 2008.

Latson, L., Sebek, B., Powell, K., "Automated Cell Nuclear Segmentation in Color Images of Hematoxylin and Eosin-Stained Breast Biopsy", *Anal. Quant. Cytol. Histol*, vol. 26, no. 5, pp. 321–331, 2003.

Malik, J., et al., "Contour and texture analysis for image segmentation", *Int. J. Comput. Vis*., vol. 43, no. 1, pp.7 - 27, 2001.

Malik, Khan, "Modified Watershed Algorithm for Segmentation of 2D Images", *Journal of Information science & Information Technology*, 6: No.3, pp. 546-552. 2009.

Mariuzzi, L., et al., "Quantitative study of ductal breast cancer progression: Signatures of nuclei in proliferating breast lesions and in situ cancers," *Advances in Clinical Pathology, 4*, 87-97, 2000.

Naik, S., et al., "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2008, pp. 284-287.

CITED LITERATURE (Continued)

Naik, S., et al., "Gland Segmentation and Computerized Gleason Grading of Prostate Histology by Integrating Low-, High level and Domain Specific Information" *Proceedings of 2nd Workshop on Microsopic Image Analysis with Applications in Biology*, 2007.

Partin, A.W., et al., "A comparison of nuclear morphometry and Gleason grade as a predictor of prognosis in stage A2 prostate cancer: a critical analysis". *J Urol*, 142(5): p. 1254-8, 1989.

Parvin B., et al., "Iterative Voting for Inference of Structural Saliency and Characterization of Subcellular Events", *IP (16)*, No. 3, March 2007, pp. 615-623.

Recky, M., Leberl, F., "Windows Detection Using K-Means in CIE-Lab Color Space," *ICPR10*, pp.356-359, 2010.

Schmitt, O., Hasse, M., "Radial symmetries based decomposition of cell clusters in binary and gray level images", *Pattern Recognition* 41(6): 1905-1923, 2008.

Sertel, O., Lozanski, G., Shana'ah, A., Gurcan, M. N., "Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading," *Signal Processing Systems*, pp.169-183, 2009.

Sertel, O., et al., "Computer-aided detection of centroblasts for follicular lymphoma grading using adaptive likelihood based cell segmentation," *IEEE Trans. on Biomedical Engineering*, vol. 57(10), pp. 2613-2616, 2010.

Sohn S, Y., and Kim, H. S., "Random effects logistic regression model for default prediction of technology credit guarantee fund," *European Journal of Operational Research* 183, pp. 472 – 478, 2007.

Tsybrvskyy, O., Berghold, A., "Primary unit for statistical analysis in morphometry: patient or cell?" Analytical Cellular Pathology 18(1999) 191-202.

Urban, D., et al., "Evaluation of biomarker modulation by fenretinide in prostate cancer patients, *Eur Urol,* 35(5-6): p. 429-38, 1999.

Veltri, R.W., et al., Prediction of prostate-specific antigen recurrence in men with long-term follow-up postprostatectomy using quantitative nuclear morphometry. Cancer Epidemiol Biomarkers Prev, 17(1): p. 102-10, 2008.

CITED LITERATURE (Continued)

Waheed, S., et al., “Computer Aided Histopathological Classification of Cancer Subtypes,” *BIBE* pp. 503-508, 2007.

Wolfe, P., et al., “Using Nuclear Morphometry to Discriminate the Tumorigenic Potential of Cells: A Comparison of Statistical Methods,” *Cancer Epidemiol Biomarkers Prev*, 13(6), pp. 976- 988, 2004.

Yang, Q., et al., "Localization of Saliency through Iterative Voting," *ICPR*, vol. 1, pp.63-66, 2004.

VITA

NAME KUSUMA BAPURE

EDUCATION Bachelor of Engineering,
Instrumentation Technology,
Visvesvaraya Technological University, 2009

Master of Science,
Electrical and Computer Engineering,
University of Illinois at Chicago, 2011

EXPERIENCE Graduate Assistant – Web Developer and Database Analyst
College of Pharmacy, UIC
Nov 2009 – July 2010

Graduate Research Assistant – Dr. Peter Gann,
Pathology Research, College of Medicine,
University of Illinois at Chicago;
Aug 2010 – July 2011