LUNG CANCER SUBTYPE CLASSIFICATION FROM WHOLE SLIDE HISTOPATHOLOGICAL IMAGES

by

DHEERAJ GANTI

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2015

To my parents.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervising professor, Dr. Junzhou Huang who inspired me to do this thesis without whom this thesis would not have been possible. His irreplaceable encouragement and supervision are the main reasons of the successful outcomes of my research. I sincerely express my gratitude to Dr. Heng Huang and Dr. Jeff (Yu) Lei for serving on my committee. I would like to thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial.

I would like to thank Jiawen Yao, Yeqing Li and other friends in my lab. I would also like to thank my friends Ankur Goyal, Lalit Kumar Naidu and Surya Narayanan Swaminathan for their constant support and encouragement.

December 1, 2015

ABSTRACT

LUNG CANCER SUBTYPE CLASSIFICATION FROM WHOLE SLIDE HISTOPATHOLOGICAL IMAGES

DHEERAJ GANTI, M.S.

The University of Texas at Arlington, 2015

Supervising Professor: Junzhou Huang

Lung Cancer is one of the most serious diseases causing death for human beings. The progression of the disease and response to treatment differs widely among patients. Thus it is very important to classify the type of tumor and also able to predict the clinical outcomes of patients. Majority of lung cancers is Non-Small Cell Lung Cancer (NSCLC) which constitutes of 84 % of all the type of lung cancers. The two major subtypes of NSCLC are Adenocarcinoma (ADC) and Squamous Cell Carcinoma (SCC). Accurate classification of the lung cancer as NSCLC and its subtype classification is very important for quick diagnosis and treatment. In this research, a quantitative framework is proposed for one of the most challenging clinical case, the subtype recognition and classification of Non-Small Cell Lung Cancer (NSCLC) as Adenocarcinoma (ADC) and Squamous Cell Carcinoma (SCC). The proposed framework made effective use of both the holistic features and topological features which are extracted from whole slide histopathology images. The local features are extracted after using vigorous cell detection and segmentation so that every individual cell is segmented from the images. Then efficient geometry and texture descriptors which

are based on the results of cell detection are used to extract the holistic features. We determined the topological properties from the labelled nuclei centroids to study into the potent of the topological features. The results of the experiments from popular classifiers show that the structure of the cells plays vital role and to differentiate between the two subtypes of NSCLC, the topological descriptors act as representative markers.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Cancer

Cancer is known as gathering of related illnesses. In a wide range of cancer, some portion of the body's cells, start to separate without halting and spread into encompassing tissues. Human body is made up of cells which are trillions in number. Tumor can begin any place within these cells. Regularly, the human cells develop and separate to form new cells as required by the body. When cells are old, new cells usually substitute them. But when cancer occurs, this process does not take place as it supposed to be. The old cells do not die and the new cells are formed without necessity. The cells keep on dividing without any restrictions and forms outgrowths in the body called tumors. These tumors are usually solid and are strong masses of tissue. Malignancies of the blood, for example, leukemia, by and large do not shape strong tumors. The cancerous tumors can spread into, or attack the tissues close to them, so these are called malignant. Furthermore, as these tumors develop, some growth cells can sever and move to distant spots in the body either through the blood or the lymph node framework and shape new carcinogenic tumors a long way from the original tumor location. But benign tumors are not like malignant tumors. They do not spread or attack the tissues surrounding them, or the tissues close to them. After removal either by surgery or by other treatment procedures, benign tumors do not grow back. This is unlike malignant tumors, which sometimes grow back after removal. Generally, benign tumors are not life threatening, except for the benign

tumors that occur in the brain. The brain benign tumors can be risky and can even be the cause of death of a person.

1.1.1 Contrariety between normal human cells and cancer causing cells in the Human body

Cancer cells vary from typical cells from various perspectives that permit them to become intrusive and unmanageable. The imperative contrast is that disease cells are not as specialized as normal human cells in the body. Normal human cells can grow into cells performing particular functions in the body but carcinogenic cells cannot. This causes them to spread widely without a halt. The body uses a mechanism called programmed cell death also called as apoptosis, where it does away with unwanted cells. The cancer cells do not listen to signals sent by the body to stop dividing. The area around tumor cells in some cases, like the non-carcinogenic human cells, blood vessels, gets affected. This area is called as microenvironment. The cancer cells can affect the normal cells in such a way that they are forced to create blood vessels to feed the tumors and can get rid of the excreta from the tumors. The immune system, a system of organs and concentrated cells, that shields the body from diseases and different conditions are frequently dodged by the cancer cells. Despite the fact that the immune system typically expels harmed, abnormal cells from the body, some cancer cells stow away from the immune framework. Some tumors even utilize runaway immune responses to escape getting rid from the body or killed.

1.2 Lung Cancer

Lung cancer is the type of cancer in which there is unchecked growth of unusual cells either in one or in both the lungs. These anomalous cells do not perform the functions of healthy human cells and do not mature into normal cells. This abnor-

mality affects the proper regular functioning of the lung of supplying oxygen to the human body through blood. Though there are many advances in treatment procedures, the lung cancer which is at an advanced stage or late stage is not often easily curable. The field of lung cancer screening is a very challenging, well-paced and interdisciplinary. It is not free from controversies either. The perfect circumstance is to look the screening to permit it to develop as a general wellbeing technique.

1.2.1 Types of Lung Cancer

The lung cancers are usually classified into based on their appearance in microscope:

1. Small Cell lung cancers
2. Non-small cell lung cancers

The treatment of these two types of lung cancers is quite different as they have varied features for growth and spread. So it is extremely important to differentiate between Small cell lung cancers and Non-small cell lung cancers [1].

1.3 The Problem and Current challenges

Non-small cell lung cancer (NSCLC), the most widely recognized sort of lung tumor, is one of genuine ailments bringing on death for human beings. Computer-aided diagnosis and survival prediction of NSCLC is of great significance in diagnosis and treatment of people suffering from lung cancer [1]. The prognosis of lung malignancy is still poor, with five-year survival rate of roughly 10% in many nations. NSCLC accounts for the majority (84%) of lung cancer. Two major types of NSCLC

[1] www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-what-is-non-small-cell-lung-cancer

are adenocarcinoma (including bronchi alveolar carcinoma) which is about 40% and squamous cell carcinoma about 25 - 30% [2, 1].

1.3.1 Adenocarcinoma

Adenocarcinoma [3] is a kind of malignancy that is formed in glands which secretes mucus or fluids in different parts of the body. It generally occurs in the following cancer forms:

1. Lung cancer- Adenocarcinoma being the most common type of Non-small cell lung cancer.
2. Prostate cancer- This is the type of cancer that occurs in the prostate glands, adenocarcinoma being the cause of 99% of such cancers.
3. Pancreatic cancer- This type of adenocarcinoma occurs in pancreatic ducts of human beings.
4. Colorectal cancer- This adenocarcinoma occurs in the intestinal glands inside colon and / or rectum being the cause of 95% of colon or rectal cancers.

1.3.1.1 Diagnosis of adenocarcinoma

Diagnosis, generally differ as per the location where the cancer occurs. The tests for diagnosis include:

1. Biopsy- The surgical removal of specimen of unusual tissue in the body. The pathologists examine the dissected tissue to identify the carcinoma. After the biopsy and it is known that the cancer is present, analysis is done to find out the cancer originated from the location it is found or elsewhere in the body.
2. Magnetic resonance imaging- MRI as it is popularly known is a noninvasive restorative test that doctors use to analyze and treat medical conditions. It

uses radio frequency pulses, magnetic fields to produce detailed cross sectional images of body structures.

3. Computerized tomography- These are X-ray tests which create cross sectional images of organs, human body using X-rays and with help of computer systems. CT scans are also done during the course of the treatment to check the extent and adequacy of current treatment.

1.3.1.2 Adenocarcinoma treatment and therapy options

Usually the treatment for adenocarcinoma is provided based on the region on the human body it grows. But generally the treatment procedures include:

1. Surgery- Adenocarcinoma is treated with removal of destructive glandular tissue, and also some tissues that surrounds it. Insignificantly obtrusive surgical treatment strategies can decrease the healing time and reduce the risk of contamination after the surgery.
2. Radiation therapy- Radiation therapy is used alongside the combination of chemotherapy and surgery. Sophisticated radiation therapy, to target adenocarcinoma tissues, takes the aid of image guidance before and after the treatment. This procedure spares the healthy tissues from being damaged due to radiation.
3. Chemotherapy-The use of drugs for the treatment of cancer cells is called chemotherapy. The drugs kill the carcinogenic cells in a specific region of the body or the entire body. Chemotherapy can also be used in combination of radiation therapy and surgery to increase the adequacy of the treatment.

1.3.2 Squamous Cell Carcinoma

Sometime ago, squamous cell carcinoma [4] were more regular than adenocarcinomas. At present squamous cell carcinomas represent about around 25 % of the

whole NSCLC cases. Otherwise called epidermoid carcinomas, squamous cell carcinomas most often times occur in the bronchi within the central region of the chest. This sort of lung tumor frequently stays inside the lung. This cancer spreads through the lymph nodes and creates cavities. Squamous cell carcinoma is usually found in the regions of the body affected by UV rays from the sun, long term exposure to chemicals such as arsenic, exposure to radiation of any form.

1.3.2.1 Symptoms

The usual symptoms of squamous cell carcinoma [2] are coughing up blood, Wheezing (high pitched whistling sound occurring while a person breathing in or breathing out), constant cough. Squamous cell carcinoma is usually found in big airways. They therefore usually show symptoms at an early stage as compared to other forms of lung cancer. They usually obstruct the airways of the lungs causing infections like pneumonia and damage part of the lung. There is a syndrome called Pancoast syndrome. The syndrome starts at the beginning of the lungs and move on to other parts of the body adjacent to them, is mostly caused by Squamous cell carcinoma. People with squamous cell carcinoma are likewise more inclined to encounter a raised calcium level (hypercalcemia) which can bring about weak muscles and other issues. Hypercalcemia is one of the side effects of paraneoplastic disorder, and is created by a tumor which secrets a hormone-like substance that brings the calcium level up in the blood.

[2] http://lungcancer.about.com/od/typesoflungcancer/a/Squamous-Cell-Carcinoma-Of-The-Lungs.htm

1.3.2.2 Diagnosis

The first step of detection of squamous cell carcinoma is through X-rays when there are any unusual abnormalities in the lungs. The other diagnostic methods are:

1. CT Scan- The computed tomography of the chest aids pathologists visualize the lungs and vessels inside them through non-invasive imaging techniques. The technique also involves injecting a dye known as contrast dye into the veins before scanning, so as to enable pathologists to clearly view the lungs.
2. PET Scan-Positron Emission Tomography test is a radiology test which is commonly used alongside other diagnosis measures like CT scan.
3. Bronchoscopy- A process in which large tube is inserted in the passage of mouth or through the nose in order to view the airways under medical supervision.

1.3.2.3 Stages of Cancer

The squamous cell carcinoma of the lungs can be classified into 4 stages of cancer. In stage 1, the cancer is present within the lung and has not yet spread to other parts and lymph nodes. In stage 2, the tumor already affected the nearby lymph nodes or is in a particular part of the bronchi. In stage 3, the lungs are already affected by the carcinoma. And in stage 4, other regions of the body get affected.

1.4 Goal of Thesis

Therefore accurate classification of Adenocarcinoma (ADC) and Squamous cell carcinoma (SCC) has become vital in lung cancer pathology and can provide assistance for customized treatment planning. As of now, pathologists settle on finding choice or make their decisions on cellular and inter-cellular level morphology. The greater part of current pathology finding is still in light of subjective instincts of

pathologists and the differing capacities of specialists could bring about huge understanding blunders or predisposition. Therefore, the necessity of design of a quantitative analysis framework arises to avoid subjectivity. Notwithstanding fast advance as of late, the fundamental challenge as far as the computational procedures is the need of investigating every single individual cell for precise diagnosis, since the separation of most sickness evaluations profoundly relies on upon the cell-level data, for example, its morphology, shape and appearance. Truth be told, most utilize instances of medicinal picture investigation require such exhaustive examination. On the other hand, this is extremely tedious. For instance, an entire slide histopathological picture might have billions of pixels, and even a region-of-interest (ROI) image contains a huge number of cells. So the goal of this thesis is to accurately classify the two subtypes of Non-small cell lung cancer as adenocarcinoma and squamous cell carcinoma for assisting pathologists in accurate diagnosis and personalized treatment.

1.5 Related Work

Some methods [5, 6, 7, 8, 9] have been proposed to discuss large-scale image search, but there still remain problems to be applied in whole slide pathological images. Since the conclusion of most illness evaluations exceedingly relies on upon the cell-level data, specialists proposed to investigate individual cells for precise analysis [1, 10]. In scientists [1] proposed three sorts of local features for quantitative examination. The features are namely geometry features, pixel intensity statistics and texture features. All sectioned lung cancer cells and geometric properties such as area and contour perimeter are considered by geometry features. The other values like mean, standard deviation in lab color space are calculated by pixel intensity statistics. Haralick grey-level co-occurrence matrix is used to quantify image grey-levels, sharpness, contrast and intensity changes in the images [11]. The framework for automatic detec-

tion and segmentation of cells from lung cancer images, thousands in number, result in half a million of images of the cells. After this, the texture features are extracted from the cell images and a large scale retrieval of cell images for each of the cell segmented is performed by the system for classification based on the category. The final classification results of images tested (test image) is obtained by the majority voting process of the classification of all the cells. Recent works in similar field have focused on individual cells and cellular structures and did not consider the internal organization and connection between the cells. The arrangement of nuclei has been demonstrated to be highly important marker in breast cancer histopathology [12]. In-order to describe the cellular structures, the graph-based topological or architectural features have been proposed which are based on nuclei centroids. To the best of my knowledge, not many works have been done to elaborately discuss the performance of topological features in Lung cancer.

1.6 Organization of Thesis

The first chapter of the thesis introduces Cancer, Lung Cancer, the problem and current challenges, and the goal of thesis. Chapter 2 of the thesis explains the extraction of Holistic features (local features) from the segmented and detected histopathological images. Chapter 3 puts light on the importance of topological features (architectural features) in cancer histopathology and their extraction. The methodology, the experiments, data used and the classifiers used to classify the Non-small cell lung cancer are explained in detail in Chapter 4. The thesis is concluded in Chapter 5 and also future work is discussed.

CHAPTER 2

LOCAL FEATURES

The local features (holistic features) consist of texture and cell level information, i.e. appearance and shape of every individual cell [13]. Haralick features [11] and Gabor wavelet texture [14] are calculated to measure the texture of the cells. In order to get the accurate cell-level information, there is a need for accurate and effective cell detection and segmentation method which extracts the cell nuclei [13]. A contour based minimum model approach for detection and segmentation of cells from microscopic images is adopted. This approach is motivated by [15]. The approach, detects contours which are not dependent on its shape, by using minimal priori information [15]. The segmentation bias is avoided with reference to the shape features and enables precise cell segmentation. The results of the segmentation of a patch which is randomly selected is shown below in Figure 2.1, comparing the minimum model approach with Otsu [16] and ISO (Isoperimetric) method [17]. The results clearly indicate the accuracy of minimum model as it can precisely detect robust boundaries of individual cells. The geometry and textural features are extracted from the cells based on the segmented boundaries of cells. The geometry properties consist of area, compactness, eccentricity, twelve other similar properties, and also Zernike shape features [13]. The texture features are then extracted from each cell. A cell which has a smooth appearance usually has not many textures. If a cell has many textures, then it appears rough and exhibit a varied pixel intensities. The Haralick [11] and Gabor "wavelet" features [14] are measures as texture properties of objects [13].

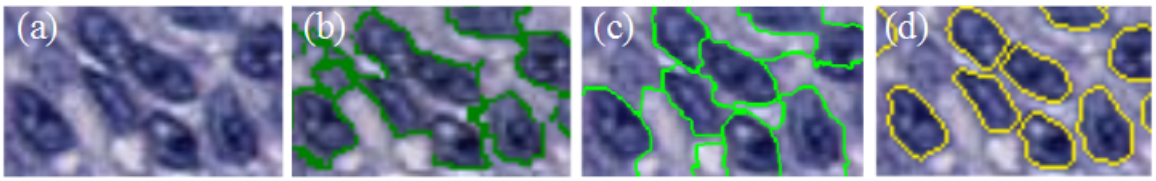

Figure 2.1. Segmentation results on three methods on arbitrarily choosen patch. From left to right: a. Original image, b. Otsu, c. ISO, d. Minimum-model.

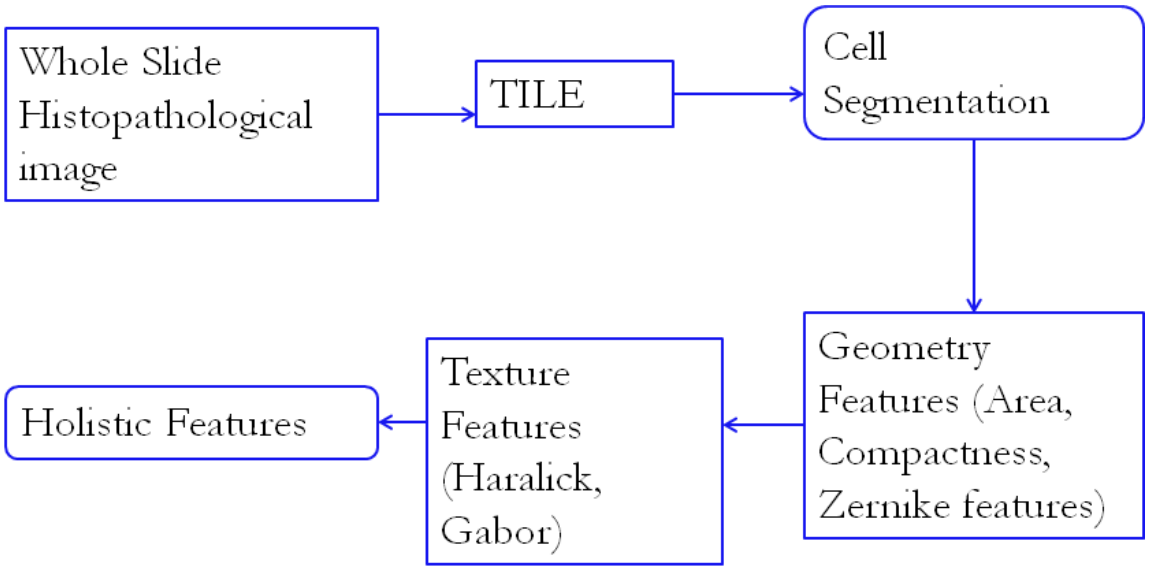

Figure 2.2. Architecture of extracting Holistic_Features.

2.1 Cellprofiler

CellProfiler [1] is free open-source software or framework which enables pathologists and biologists who do not have specific knowledge in fields like computer vision, image processing and programming for quantitative measurement and evaluation of

[1] http://www.cellprofiler.org/

the phenotypes from images, thousands in numbers. The process is automatic or involves less human interaction or supervision.

In the experiments, the holistic features like geometry features, texture features and histopathological image processing is done using the CellProfiler [18].

2.2 Histopathological Image processing

The images from the tile (ROI) extracted from the whole slide image is taken and different image processing techniques are applied on them.

2.2.1 Unmix colors

This produces separate images for histological stained images as per the dye stain. Different gray scale images are formed from given color images which are stained by using dyes which are light absorbing. Dyes are accepted to ingest a measure of light in the RGB (red, green and blue) channels that is incremented proportionally with increase in the amount of stain. This module produces gray scale images by differentiating two or more than two stains from the background. In cellprofiler, there are different default combinations of dyes. The user can also use the custom mode feature to use images stained with a single dye and compare or calibrate two images after the separate staining. In this experiment, we use Hematoxylin and Eosin for staining the lung biopsy tissue (Fig:2.3). The stain Hematoxylin is specifically used to stain nucleic acids and endoplasmic reticulum generally. And the stain Eosin is used to stain elastic, collagen and reticular fibres.

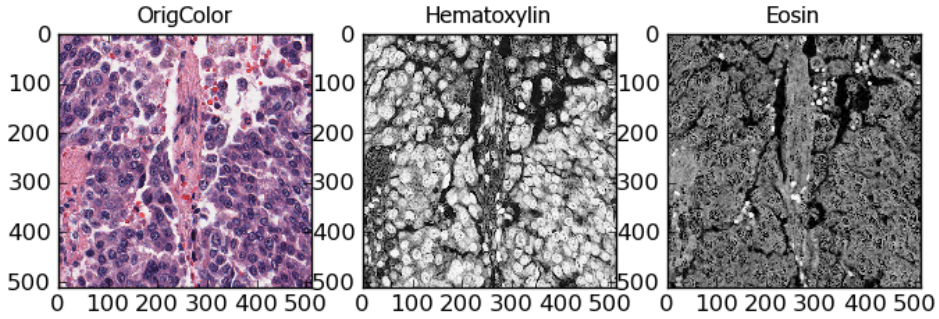

Figure 2.3. Unmix colors.

2.2.2 ColorToGray

This module is used to convert an image with the three color channels namely red, green and blue to either one or three images of gray scale (Fig:2.4). There are two procedures to convert from color to gray scale. Those are:

1. Split- Separate gray scale images are formed after three color channels (RGB) are split.
2. Combine- A single gray scale image is formed by combining together the three channels (RGB).

We use the combining method to form gray scale image. The relative weights of all the three channels are kept equal so that they have equal weight of contribution towards the final image formed by combining process.

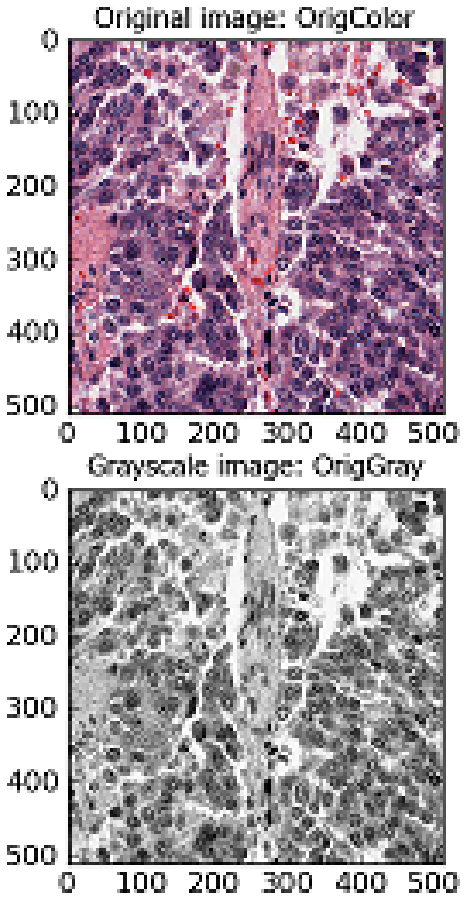

Figure 2.4. ColorToGray image.

2.2.3 ImageMath

By using this module, we perform basic calculations and operations on the intensities of the images. We can perform basic additions, subtractions, multiplications, divisions, averages , log transform, scaling of individual image intensities by a constant, inversion, two or more than two intensities of the images. We basically use this module to invert the image (Fig:2.5).

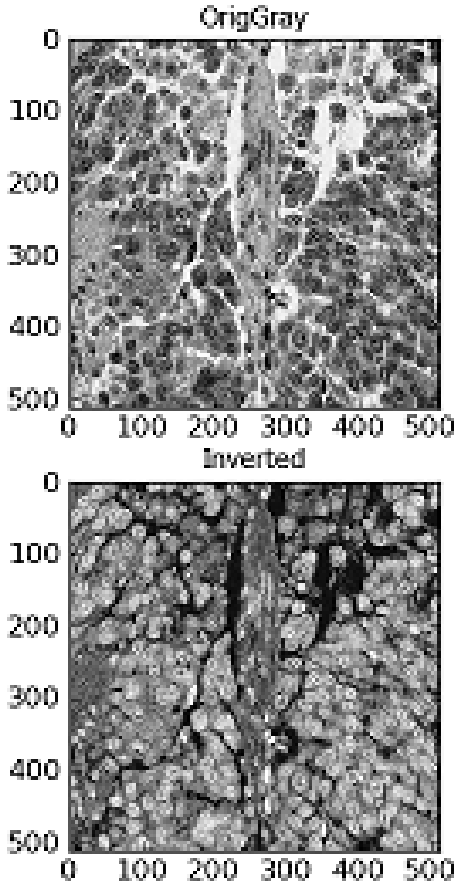

Figure 2.5. ImageMath_OrigtoInv image.

2.2.4 IdentifyPrimaryObjects

Identify primary objects module is used to distinguish essential items within the gray scale images. The images are identified as bright objects on a dim or dark background. The objects which can be found in image without the need of comparing or correspondence to a reference image are called the primary objects. Whereas the objects found with reference to previously obtained primary objects are called the secondary objects.

1. In the first cycle of IdentifyPrimaryObjects, we identify Tissues by providing inverted images as input (Fig:2.6).

2. In the second cycle, we identify HemaSubRegion from the input image stained with Hematoxylin (Fig:2.7).
3. In the third cycle of the IdentifyPrimaryObjects module, we identify EosinSubRegion from the input images stained with Eosin (Fig:2.8).

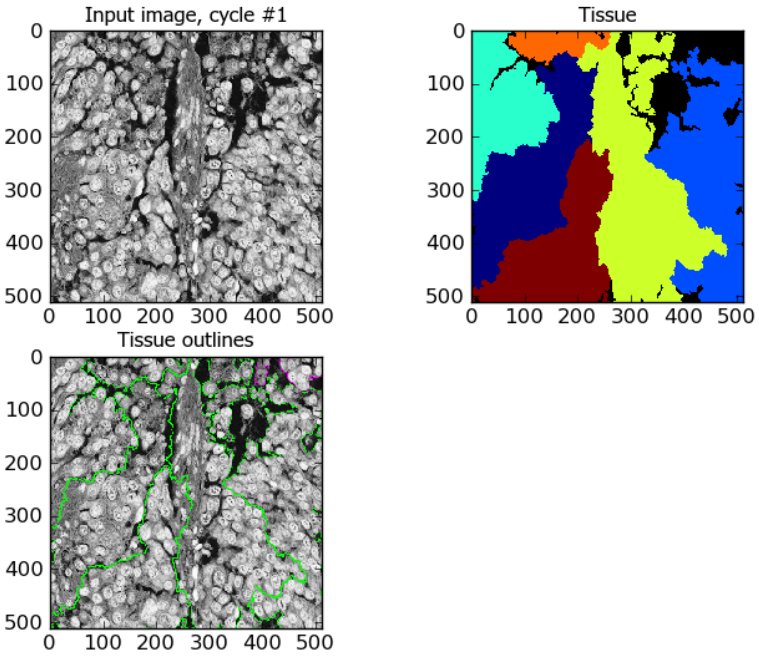

Figure 2.6. Identifyprimaryobjects image.

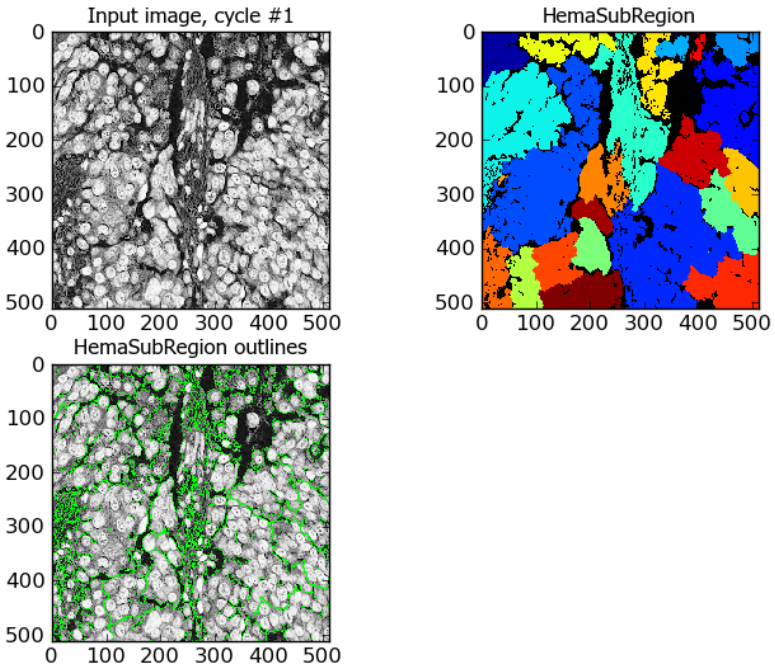

Figure 2.7. Identifyprimaryobjects(HemaSubRegion) image.

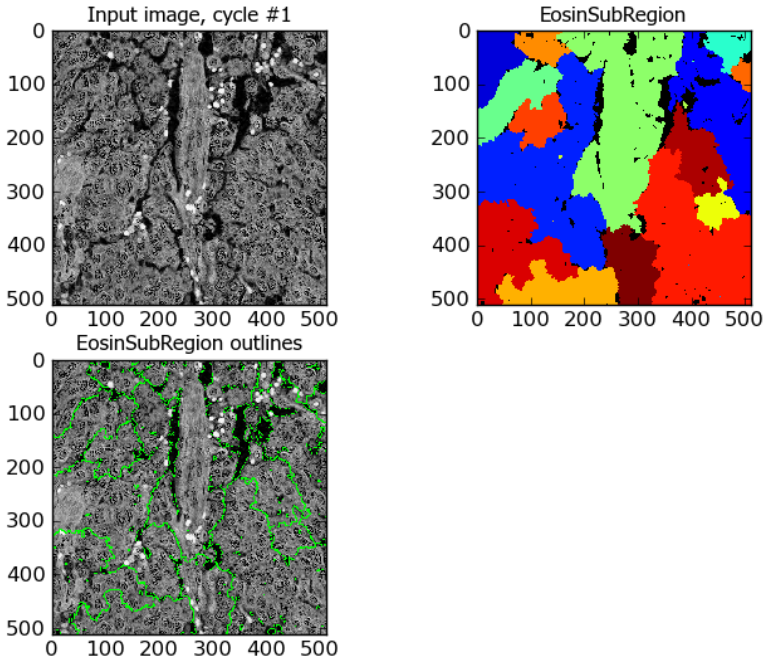

Figure 2.8. Identifyprimaryobjects(EosinSubRegion) image.

2.2.5 MaskImage

The MaskImage module is used to mask or hide, as the name implies, some parts of the image. The masking is done by using previously identified images or objects as reference. They are not considered in the pipeline for subsequent masking. In this experiment, we mask the Eosin sub regions from the primary objects in the first MaskImage module that we use (Fig:2.9). And in the second MaskImage module that we use in the experiment, HemaSubRegions are masked from primary objects by using masking objects instead of masking image (Fig:2.10).

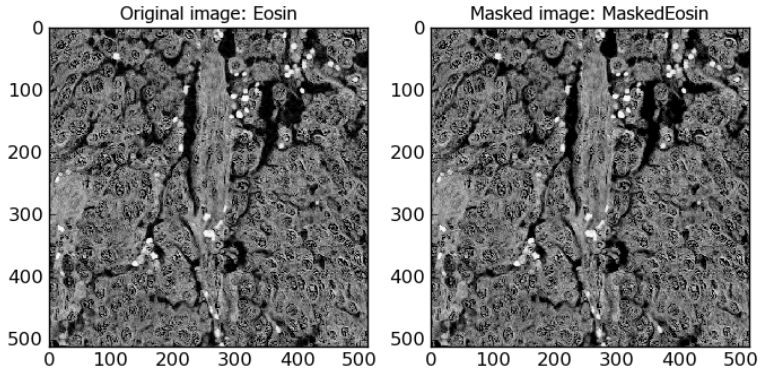

Figure 2.9. Masked_Eosin image.

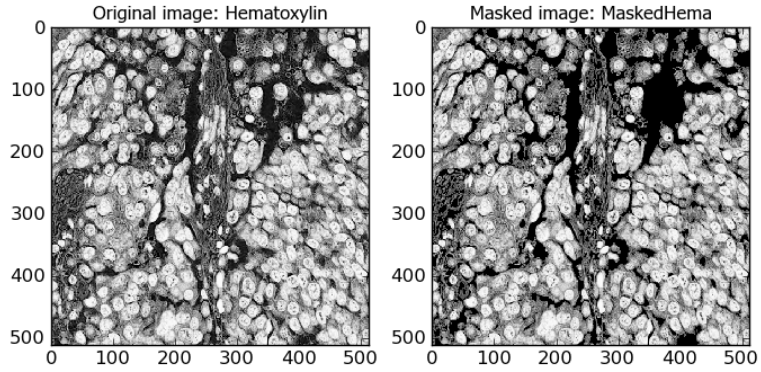

Figure 2.10. Masked_Hematoxylin image.

2.3 Detection and Segmentation

2.3.1 SMILEIdentify

Image Segmentation [19] is the procedure of dividing a digital image into different fragments. The nuclei segmentation in histopathological images is very important to simplify the image to make it easier to analyze and understand.The Nuclei detection and segmentation is not so accurate using the Cellprofiler for our process. A more vigorous or robust cell segmentation method is necessary for the purpose and we followed a minimum-model approach for the cell detection and segmentation [15] (Fig:2.11,Fig:2.12).The benefits of this approach are

1. It avoids a segmentation bias with respect to the shape features.
2. This strategy takes into consideration a precise division of a wide range of ordinary and infection related morphological features without the requirement of prior training as it uses minimal priori information and detects contours independent of their shape.

The six steps involved in the minimum-model approach [15] are

1. The detection of every plausible closed contours
2. Evaluation of Contours
3. The Generation of segmentation which is not overlapping
4. Optimization of the contours
5. Separating Concave objects from the segmentation images
6. Classification into cell nuclei and the rest objects

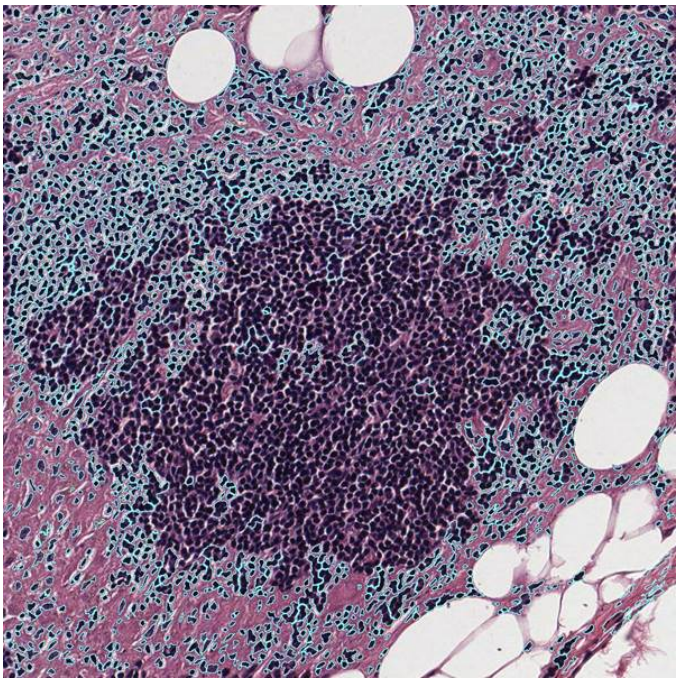

Figure 2.11. CellProfiler Segmentation image.

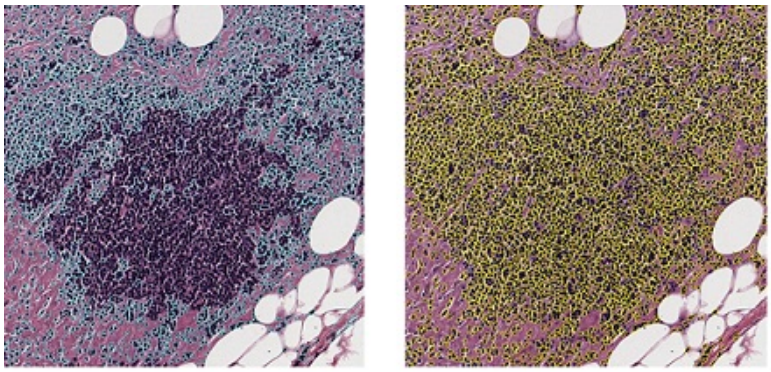

Figure 2.12. Comparison of Segmentations image.

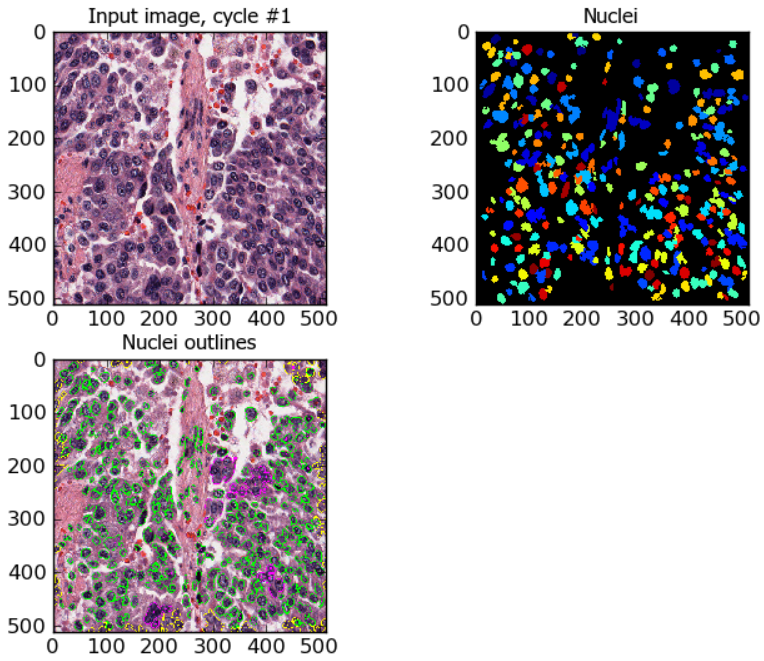

Figure 2.13. SMILE_identify image.

Accurate detection and segmentation of Nuclei from the histopathological images is done because holistic features are extracted after the process and they depend upon precise and robust segmentation. The SMILEIdentify module motivated by the minimum-model approach [15] is integrated into the CellProfiler framework (Fig:2.14) for the robust detection and segmentation for better results in our process. After the detection and segmentation, the geometry and texture features are extracted from the histopathological images, i.e. the extraction of holistic features.

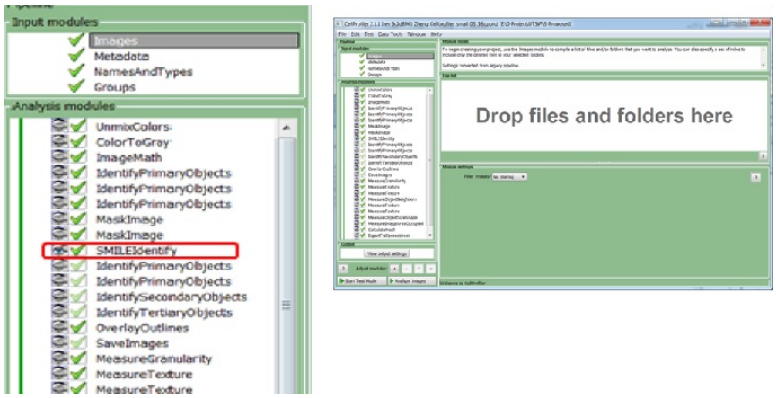

Figure 2.14. Cellprofiler_integration image.

2.4 Geometry Features

2.4.1 Overlay Outlines

This module is used to create outlines on a desired image. This creates outlines by an Identify module in a specific and desired format (Fig:2.15). The desired image can be color or gray scale and it can even be a blank image. Specific settings for this module are used in the experiment. The display images are not outlined on a blank image but from a given input image i.e. the OrigColor image. The outline display mode is set to Color which displays the outline contours around the desired image. Though the color outlines occupy more memory, they are chosen for the experiment because they create a very clear display for images having high intensity for the edges of the cell. The width of the outline is the width in pixels that can be shown on the image.

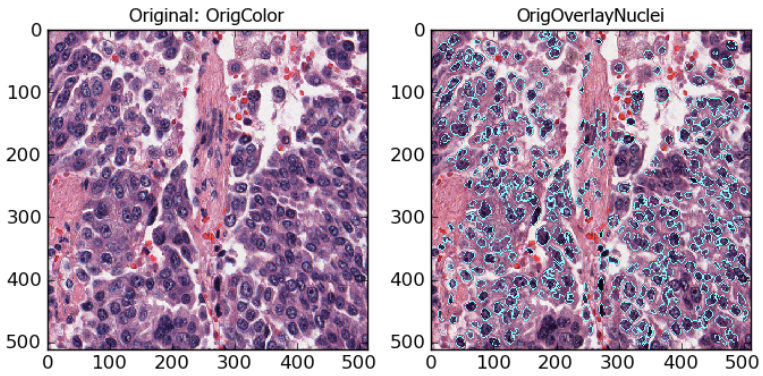

Figure 2.15. Overlay_outlines image.

2.4.2 MeasureObjectNeighbors

This module calculates the number of nearest neighbors of each given object and puts some light on the relationship between object neighbors like by what value the edge's pixels of an object are overlapping or touching its neighbor.

1. Numberofneighbors - Neighboring objects.
2. PercentTouching - by what value the edge's pixels of an object touches its neighbors.

2.4.3 MeasureObjectSizeShape

This module is used to extract shape features and size of each image given with identified objects like nuclei or cells. The objects touching the edges are removed because this module can be trusted when the objects are completely bound within the images.

2.4.4 MeasureImageAreaOccupied

This module is used to measure the total area occupied by the objects in a given image.

2.5 Texture Features

2.5.1 Haralick features

Haralick features provide information on how the intensities of pixels in images of a certain position are related with the intensities of neighboring pixels. These features originated from the co-occurrence matrix. In the Cellprofiler [18] used to measure the local features, there is a method called MeasureTexture. This method can obtain textures of images at different scales. The scale can be chosen by the user. The co-occurrence matrix thus constructed is a result of the scale chosen by the user. If we chose a scale of 2, then each pixel from the image is compared to the pixel that is present at two pixels at the right side of the given pixel. This method usually quantizes the given image into 8×8 co-occurrence matrix. It looks for the number of pixels and neighbors each having the combinations of 8×8 intensity. Using this method, around 13 features are calculated. These features are calculated on the co-occurrence matrix by making certain calculations. The features calculated using this method are angular second moment, correlation, contrast, variation (sum of squares), inverse difference moment, sum variance, sum average, entropy, sum entropy, difference entropy, difference variance, information measure of correlation1 and correlation2.

2.5.2 Gabor wavelet features

The Gabor wavelet features are the result of applying Gabor filters to images. These features are similar to wavelet features. The frequency measure in different orientations of images is measured using the Gabor features. Though they are similar to and work like wavelets in this context, these Gabor features are not wavelets by the strict definition of mathematics. These features detect bands of intensities that are correlated.

2.5.3 MeasureGranularity

The module is used generally to try series of structure elements that outputs a number of measures. It defines whether the structure elements fit in the texture of given images.

2.6 Summary of Holistic Features

The number of geometry features used are 45, texture features used are 106 in number and the pixel intensity statistics used are 4. The total number of holistic features used for the analysis used are 155.

Table 2.1. The holistic features used for analysis

Feature Set	Description	Number
f_G	Area, Perimeter, Compactness, FormFactor, Solidity, Extent EulerNumber, Eccentricity, Axis Length (Major, Minor) Radius (Max, Mean, Median), Feret diameter (Max, Min) Zernike features	45
f_T	Gabor and Haralick Features (Nuclei, Intensity)	106
f_S	Area, Perimeter, Number (Nuclei), Stain area	4
f_L	Holistic features	155

CHAPTER 3

TOPOLOGICAL FEATURES

Given a histological image, the arrangement of nuclei within the region is related to the type of cancer and this architecture can be described by graph-based techniques such as Voronoi Diagram, Delaunay Triangulation and Minimum Spanning Tree [12]. We first define the undirected and complete graph as $\mathcal{G} = (\mathcal{O}, \mathcal{E}, \mathcal{W})$ where $\mathcal{O}$ is the collection of nuclear centroids, $\mathcal{E}$ is the collection of edges connecting the nuclear centroids, and $\mathcal{W}$ is the collection of weights of each $\mathcal{E}$ [13].

3.1 Voronoi diagram

A Voronoi diagram, according to arithmetic, is dividing of a plane into areas in view of distances from a point in a particular subset of the plane. The arrangement of points is determined heretofore. For every seed, there is a comparing locale. The locale comprises of all points nearer the seed as compared to any other. These areas are known as Voronoi cells. The Voronoi diagrams are made by taking sets of points that are near one another. After taking the set of points, a line is drawn in the middle, which is at equal distance from them. The line is drawn in such a way that it is perpendicular to the line connecting the points. Thus, all points in the given Voronoi diagram [1] will be equidistant to closest two, three or more source points [20].The voronoi diagrams finds its applications in various fields like science but it has also found its way through the field of arts. The Voronoi Diagram $\mathcal{G}_{\mathcal{V}}$ includes a set of polygons $\mathcal{P} = \{P_1, P_2, ..., P_m\}$encompassing every nuclear centroids O. Every

[1]https://en.wikipedia.org/wiki/Voronoi_diagram

pixel is connected with the closest centroid and added to the related polygon P. After the construction of the graph $\mathcal{G}_{\mathcal{V}}$, the mean, standard deviation, minimum / maximum ratio, and disorder (standard deviation / mean) are calculated for the length of the perimeter, the area and length of the chord over all the polygons [13].

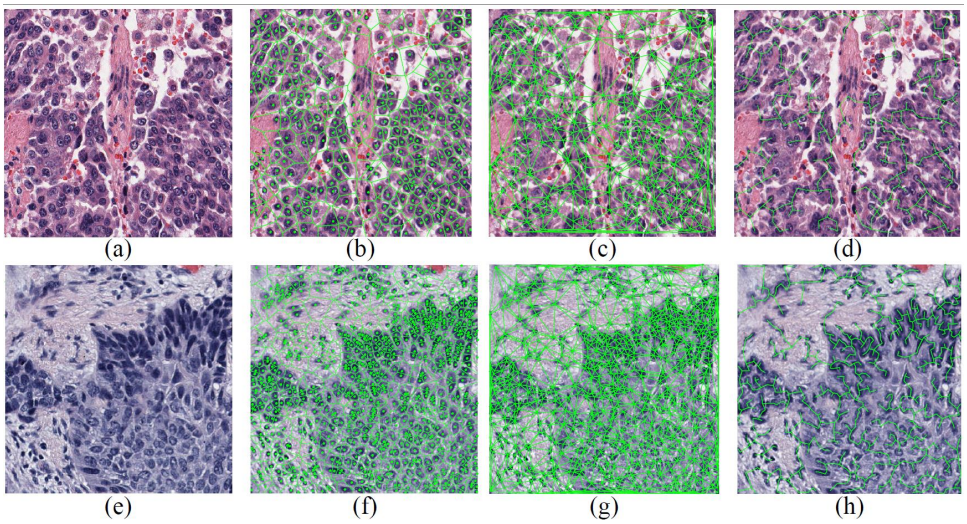

Figure 3.1. Graph-based topological features on images from two types of tumor adenocarcinoma (a) and squamous cell carcinoma (e). Voronoi Diagram (b-f), Delaunay Triangulation (c-g) and Minimum Spanning Tree (d-h)..

3.2 Delaunay triangulation

The Delaunay graph $\mathcal{G}_{\mathcal{D}}$ is a spanning subgraph of $\mathcal{G}$ and the dual graph of $\mathcal{G}_{\mathcal{V}}$. The construction is as follows: if $\mathcal{P}_i$,$\mathcal{P}_j \in \mathcal{P}$ share a side (i,j $\in$ 1,2,...,L), their nuclear centroids $o_i, o_j \in \mathcal{O}$ are connected by an edge (o_i, o_j). Following similar calculations, we calculate the standard deviation, mean, the ratio min/max and disorder for side length and area of all the triangles in the Delaunay graph $\mathcal{G}_{\mathcal{D}}$ [13]. The Delaunay triangulation was named after its creator, a mathematician from Russia, Boris Nikolaevich Delaunay. Delaunay triangulations amplify the base point of the considerable angles of triangles which are in triangulation. The Delaunay triangulation, have a

tendency to dodge thin triangles. The Delaunay triangulation of a distinct point set $\mathcal{P}$, by and large position relates to the dual graph of the Voronoi outline for $\mathcal{P}$. Uncommon cases incorporate the presence of three points on a single line and also the presence of four points on a circle. Put in simple terms, the Delaunay triangulation is a very proficient approach to draw and form triangles between set of points. In this method, every point is joined by lines to its nearest neighbors, in a manner where all the lines form triangles and generally they do not meet. The surface is thus totally secured with a pleasant layer forming distinctive triangular tiles. The Delaunay triangulation has significant utilizations in varied fields of computer science and graphics.

3.3 Minimum Spanning Tree

According to [13], an edge weighted graph $\mathcal{G}$ is one where each edge is associated with weights or costs. Every single edge can likewise be relegated with a weight. The weight is a number speaking to how unfavorable it is, and utilize this to dole out a weight to spanning tree by calculating the summation of weights of the number of edges in that particular spanning tree. A minimum spanning tree of such edge weighted graph is one whose weight, i.e. the sum of weights of edges in not larger than weight of any other spanning tree [21]. Putting simply, the minimum spanning tree $\mathcal{G}_{MST}$ connects together all the vertices in the undirected graph with minimum total weight of edges. One undirected graph may have various spanning trees.

3.4 Nuclear Features

For any nuclear centroid $o_i \in \mathcal{O}$, a realted nuclear neighborhood is defined as $\eta^{\zeta}(o_i) = \{o_j :\| o_i - o_j \|_2 < \zeta, o_j \in \mathcal{O}, o_j \neq o_i\}$, where $\zeta \in \{10, 20, ..., 50\}$. Also,

we roughly calculate the minimum radiusζ^*such that $\mid \eta^{\zeta^*}(o_i) \mid \in 3, 5, 7$ and also determine the standard deviation, disorder and mean over all the nuclear centroids [13].

3.5 Summary of Topological Features Used

The total number of topological features used for the analysis are 48, out of which features from Voronoi diagram, Delaunay triangulation, Minimum spanning tree and Nuclear features are 12, 8, 4 and 24 respectively.

Table 3.1. The topological features used for analysis

Feature Set	Description	Number
f_V	Polygon area, perimeter and chord length (mean, std dev., min/max ratio, disorder)	12
f_D	Triangle side length and triangle area (mean, std dev., min/max ratio, disorder)	8
f_{MST}	Edge length (mean, std dev., min/max ratio, disorder)	4
f_{NF}	Distance to {3,5,7} nearest nuclei: mean, std dev., disorder Nuclei in {10, 20,...,50} pixel radius: mean, std dev., disorder	24
f_{Arch}	Topological Features	48

CHAPTER 4

METHODOLOGY

We take the slides stained with Hematoxylin and Eosin. The stained tissues of biopsy from lungs are then scanned after magnifying them at 40×. The actual slides are very high resolution in nature. So certain areas of the slide are extracted from the desired regions. The extraction is based on expert opinions. Skilled pathologist's advices on certain regions of the tissue that are labelled are taken. And from these regions, many image tiles are obtained. These retain the local and holistic features. Thus by retaining these features in the histopathological images, we can receive an advantage of precise and correct analysis.

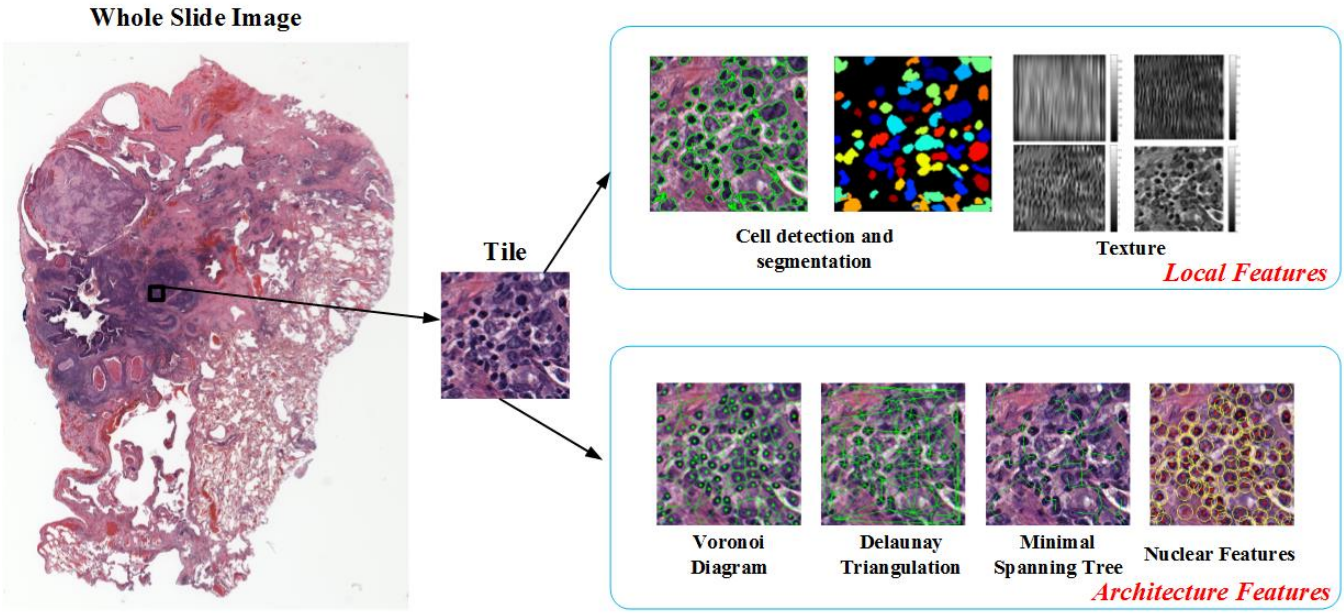

Figure 4.1. Overview_Framework.

4.1 Summary of holistic and topological features

Table. 4.1 lists both holistic and topological features used in this paper. In summary, 155 holistic (local) features include geometry and texture properties of individual cells and the whole image tile. 48 topological (architecture) features are calculated based on three kinds of graph and nuclear neighborhoods.

Table 4.1. The total features used for analysis

Feature Set	Description	Number
f_G	Area, Perimeter, Compactness, FormFactor, Solidity, Extent EulerNumber, Eccentricity, Axis Length (Major, Minor) Radius (Max, Mean, Median), Feret diameter (Max, Min) Zernike features	45
f_T	Gabor and Haralick Features (Nuclei, Intensity)	106
f_S	Area, Perimeter, Number (Nuclei), Stain area	4
f_L	Holistic features	155
f_V	Polygon area, perimeter and chord length (mean, std dev., min/max ratio, disorder)	12
f_D	Triangle side length and triangle area (mean, std dev., min/max ratio, disorder)	8
f_{MST}	Edge length (mean, std dev., min/max ratio, disorder)	4
f_{NF}	Distance to {3,5,7} nearest nuclei: mean, std dev., disorder Nuclei in {10, 20,...,50} pixel radius: mean, std dev., disorder	24
f_{Arch}	Topological Features	48
f	Holistic and Topological Features	203

4.2 Experiments

After carefully selecting and extracting the required regions by following the pathologist's advise, 100 slides of the images from the National Lung Screening Trial (NLST)are taken. Out of these slide images, the adenocarcinoma (ADC) are 50 in number and Squamous cell carcinoma (SCC) tissue slides are another 50 in number.

The required local features are calculated using the Cellprofiler[18]. The Cellprofiler is a free and open source software. The software is created so as to make it feasible and easy for researchers and analysts to measure phenotypes and local features from hundreds and thousands of images with little direct human control and get precise results as well.

4.2.1 Data

The images from NLST (National Lung Cancer Screening Trial) [22] data portal have been taken for experiments. The images of adenocarcinoma and squamous cell carcinoma lung cancer were used from the data portal.

4.2.2 Experimental setup

1. Cellprofiler - For obtaining local features.
2. Matlab (2013b) - For obtaining topological features.
3. R studio (Ver.0.98.1103) - For classification and analysis of the cancer and non-cancer and also the subtypes- adenocarcinoma and squamous cell carcinoma. The classification is mainly based on the local features and topological features.
4. System used with configurations - 3.4GHz Intel core i7 4770 CPU with a RAM of memory size 16 GB.

4.2.3 Data Specifications

The data is prepared using histopathological image slides obtained from NLST (National Lung Cancer Screening Trial). The data is made using 100 histopathological image slides. These image slides that were used for the data contain 50 adenocarcinoma (ADC) and 50 squamous cell carcinoma (SCC) tissues [13].

4.2.4 Results - ROC curves

The ROC curve is one of the fundamental tool for the assessment of the diagnostic tests [1].The ROC curve helps in the creation of a full and detailed sensitivity-specificity report. The true positive rate or sensitivity is plotted against the function of the false positive rate or specificity. The sensitivity [23] is the probability of a test outcome being positive when they are truly positive and specificity [23] is the probability of the test outcomes being negative truly as such.

1. Sensitivity or True Positive Rate (TPR) = TP/P = TP / (TP + FN)
2. Specificity or True Negative Rate (TNR) = TN/N = TN / (TN + FP)

   Where TP is True Postive, FN is False Negative, TN denotes True Negative, FP is False Positive, P is Positive and N is Negative.

4.3 Classifiers

We used different classifiers in order to compare the scheme. Cross Validation, is a model approval procedure for evaluating how the consequences of a factual investigation will sum up to independent information set [24]. We use a method called k-fold cross validation. And we here used k=10 i.e., 10 fold cross validation. Of the 10 subsamples, a solitary subsample is held as the approval information (validation data) for testing the model, and the other 9 subsamples are utilized as preparing information (training data). The cross-validation process is then rehashed 10 times (the number of folds), with each of the 10 utilized precisely once as the approval information. The 10 results from the folds can then be found the middle value of (or generally consolidated) to create a solitary estimation. The benefit of this technique over repeated random sub-sampling is that all observations are utilized for both preparing

[1]https://www.medcalc.org/manual/roc-curves.php

and validation, and every perception is utilized for acceptance (validation) precisely once [24]. We use KNN, Random Forest, SVM, PLR, Navie Bayes for classification.

4.3.1 KNN

KNN is a straightforward algorithm that stores every single accessible case and characterizes new cases taking into account a similarity or closeness measure [25]. KNN is a non-parametric lazy learning algorithm [26]. By non-parametric technique, it implies that that it does not make any presumptions on the fundamental information appropriation. This is really valuable, as in this present reality, the greater part of the reasonable information does not comply with the regular hypothetical presumptions made. Non parametric calculations like KNN act the hero here. Being a lazy algorithm, Knn does not use training data points for generalization [26]. This implies there is very minimal training and the phase of training is very quick. Absence of generalization implies that KNN keeps all the preparation information (data for training). All the more precisely, all the preparation information is required amid the phase of testing. We used k-nearest neighbor method as the baseline classification method because of its simplicity and its efficacy [27].

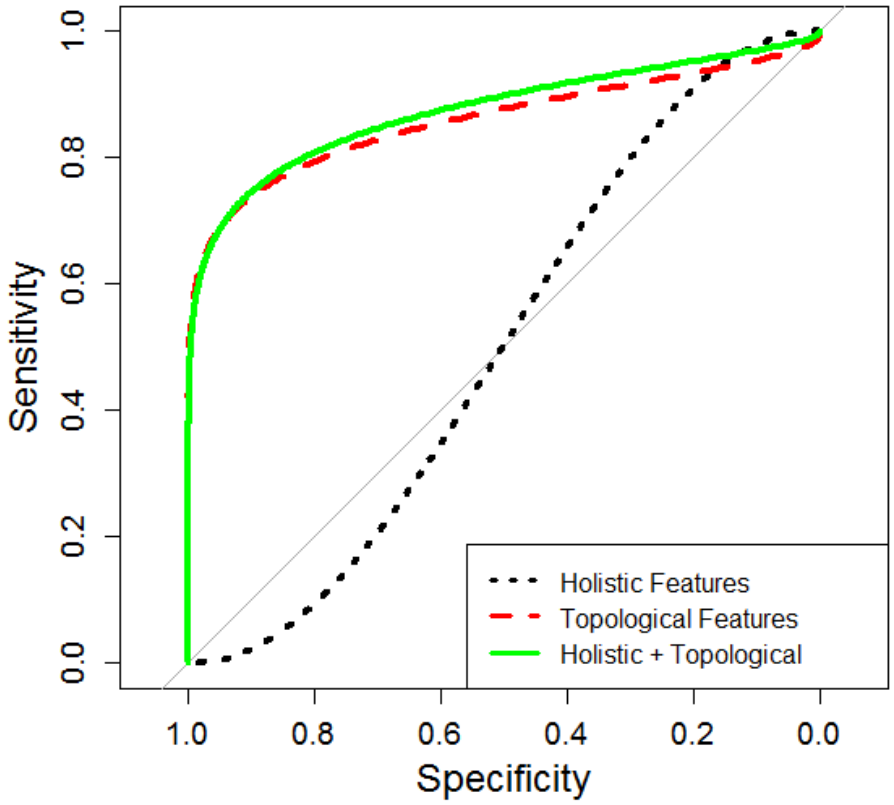

Figure 4.2. Knn_ROC.

(Area Under Curve: Holistic: 0.5025, Topological: 0.8595, Holistic + Topological: 0.8749)

From the figure, we can see that the ROC curve for combination of holistic and topological features is higher and better as compared to the curves for holistic and topological features separately. The Area under the curve also proves the same for the combined use of holistic and topological features when compared to holistic and topological features individually.

4.3.2 Random Forest Method

Random forest creates a number of decision tress. The decision trees are created depending on the random selection of data and also the selection of variables randomly. The class of the dependent variable is determined by the class based on many decision trees. The major beliefs of random forest algorithm being most of the decision trees in the random forest predict the correct classes for most of the given data and the trees do not predict correctly at different places. The voting for each observation can be done and the class of the observation can be determined on the basis of the results of the voting. This voting and classification is considered to be much nearer to the exact classification. The ensemble learning method Random Forest (RF)[28], ranks the importance of features in classification using the permutation scores for each rank. These ranks are generated in Random Forest method.

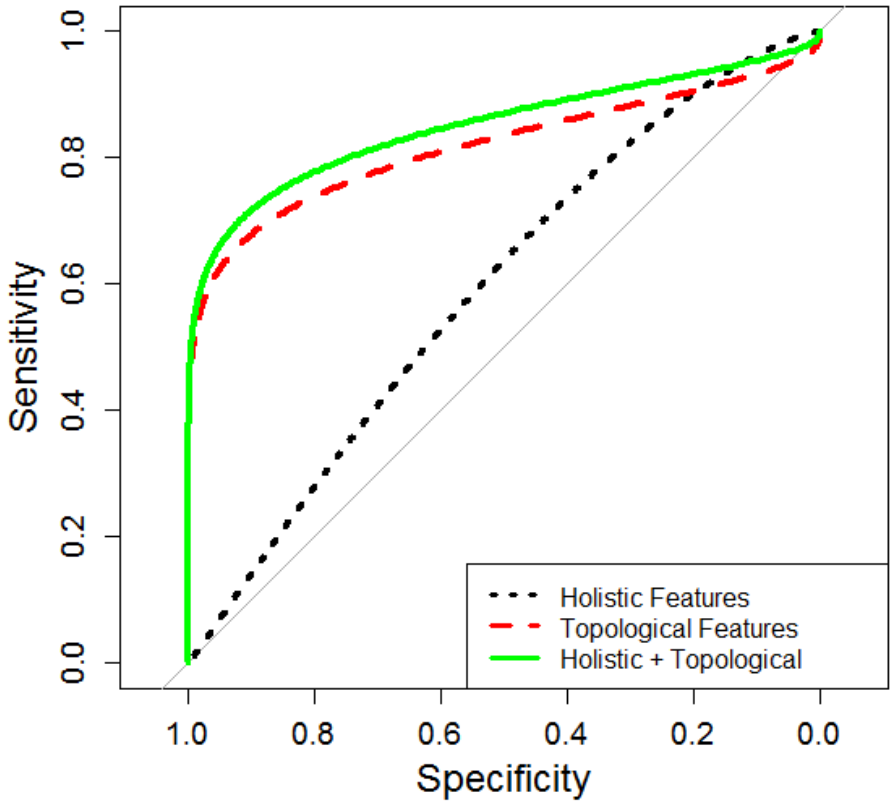

Figure 4.3. RF_ROC.

(Area Under Curve Holistic : 0.5932, Topological: 0.818, Holistic + Topological: 0.8501)

The figure shows that the ROC curves for the combined use of topological and holistic features has greater accuracy than each one of them taken individually. The Area under the curve also gives a similar conclusion because the AUC is greater for the holistic and topological features put together.

4.3.3 Support Vector Machines

The Support vector machines [2] are learning models that are supervised in nature. The SVM models are models with related learning algorithms. These learning models perceive examples, patterns, and investigate data given. The SVM models are generally used in classification and regression models. The SVM model works by training the examples or data to fall into either one of the two categories. Then the model is built using the SVM algorithm and the new entry is made to fall into either of the two categories. In simple terms, SVM model is portrayal of the samples as points in space. The points are mapped so that the illustrations of the different classes are isolated by a reasonable crevice that is as wide as could be expected under the circumstances. The new entries or cases are then represented in the same space as the previous trained examples. The new cases are then anticipated to have a place with a classification in light of which side of the crevice they fall on.We used Support Vector Machine as it is widely used for breast and prostate cancer diagnosis [29].An RBF kernel with optimized gamma value for SVM is chosen for the experiments [13].

[2]https://en.wikipedia.org/wiki/Support_vector_machine

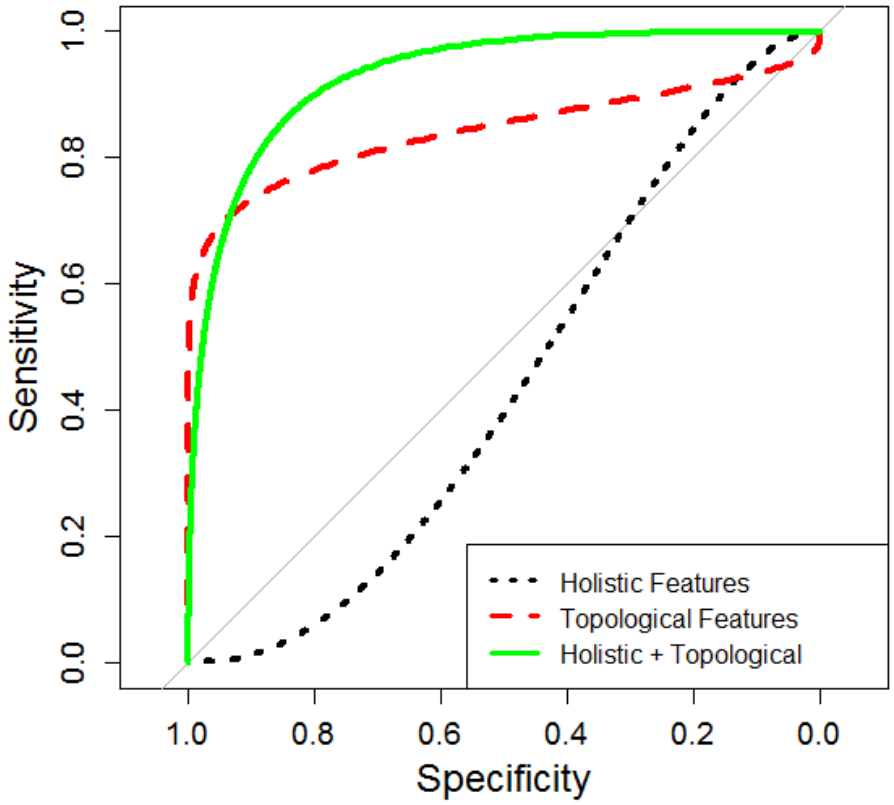

Figure 4.4. SVM_ROC .

(Area Under Curve Holistic : 0.4422, Topological: 0.843, Holistic + Topological: 0.931)

The ROC curves from the above figure speaks for the accuracy of the integration of holistic and topological features and the Area under the curve provides us with similar evidence showing the combined topological and holistic features has greater efficacy than individual features from holistic and topological.

4.3.4 Penalized Logistic Regression

The PLR is an efficient lasso regularization path for logistic regression [30].The Penalized Logistic Regression (PLR) is a classification and regression method that effectively enables selection of features from data which is high dimensional in nature.

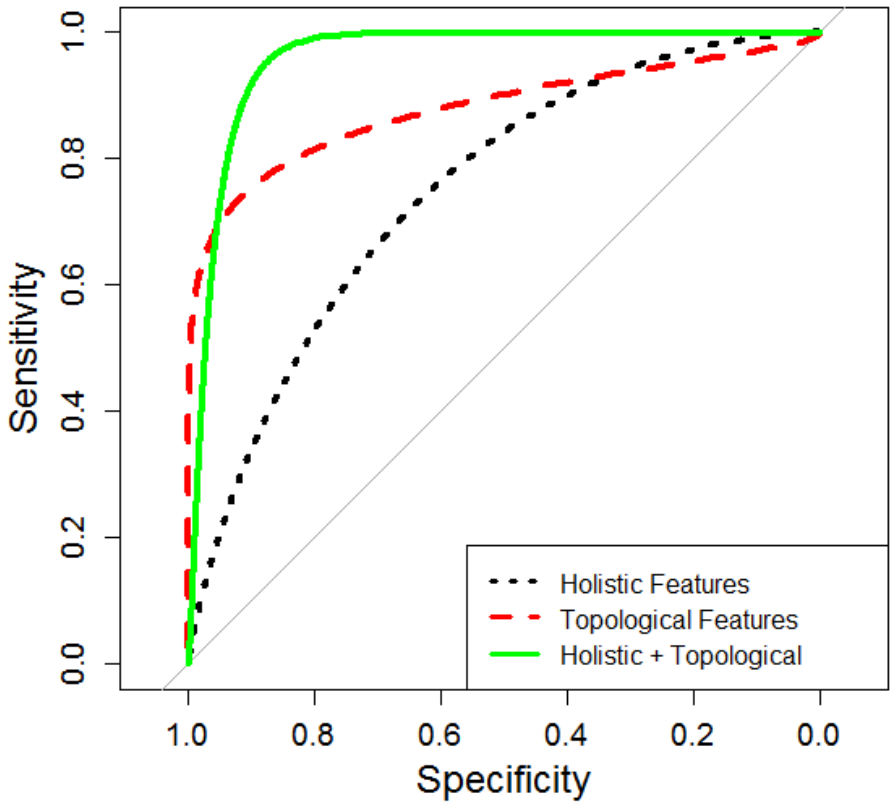

Figure 4.5. PLR_ROC .

(Area Under Curve Holistic : 0.7497, Topological: 0.8791, Holistic + Topological: 0.9596)

The ROC curves and the Area Under the Curve (AUC) show that the combined use of holistic and topological features works better in our experiments and has a

greater accuracy as compared to separately considering the holistic and topological features.

4.3.5 Naive Bayes

The Bayesian Classification speaks to an learning technique which is supervised and additionally measurable system for characterization and classification. Expects a basic probabilistic model and it permits us to catch instability about the model principled by deciding probabilities of the results. It can take care of analytic and prescient or predictive issues [31]. The classification is known after Thomas Bayes (1702 - 1761).

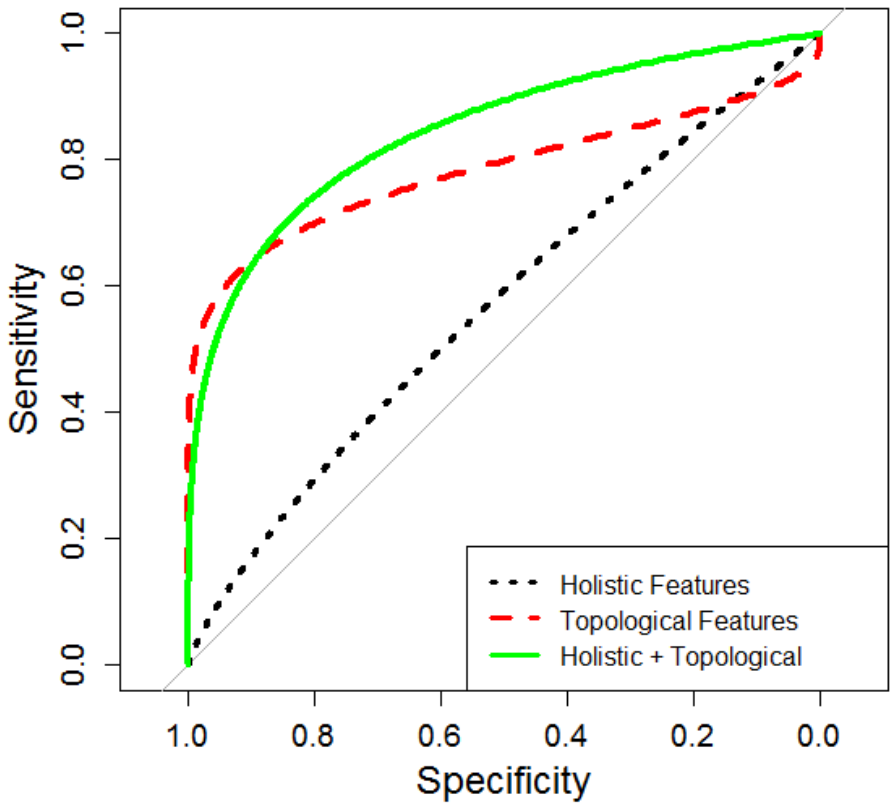

Figure 4.6. NB_ROC .

(Area Under Curve Holistic : 0.5399, Topological: 0.8924, Holistic + Topological: 0.8887)

The ROC curve shows when we use both holistic and topological features, the accuracy is better than when we use them separately. But in some cases in Naive Baiyes in our experiments, the AUC for topological is better than the combination of holistic and topological features taken together.

4.4 Results on multiple random split

The classification accuracy table shows that the combined use of holistic and topological features has a greater accuracy and efficiency as compared to only using either holistic or topological features separately for classification though topological features have better accuracy than holistic features in most cases. This proves the importance of topological features and the combined use of holistic and topological features.

Table 4.2. Evaluation of the classification accuracy. From left to right: holistic features, topological features, both holistic and topological features

	Mean	STD	Mean	STD	Mean	STD
kNN	0.4938	0.0411	**0.7969**	0.0397	0.7625	0.0493
RF	0.5250	0.0655	0.8031	0.0467	**0.8187**	0.0384
SVM	0.5094	0.0860	0.7969	0.0630	**0.8094**	0.0311
PLR	0.6531	0.0498	0.8125	0.0607	**0.8375**	0.0527
NB	0.5312	0.0625	0.8064	0.0484	**0.8468**	0.0615

4.5 Conclusion

In this chapter, we investigated the accuracy and importance of topological features in lung cancer pathology.Since not many works have been done on the classification of Non-Small Cell Lung Cancer, the principal contribution of this thesis is to enable cellular-level analysis by using holistic features. The results from experiments demonstrate that topological features are really effective to distinguish two types of Non-small cell lung cancer as adenocarcinoma and squamous cell carcinoma. Though topological features provide important information for classification, they heavily depend on the detection accuracy of the nuclei centroids. In our future work, we plan to investigate performances when nuclei centroids are given by automated methods, e.g. Deep Learning [32, 33].

CHAPTER 5

CONCLUSION AND FUTURE WORK

This research was aimed at subtype classification and diagnosis of Non-small cell lung cancer. According to [1, 10, 13] , the diagnosis of most disease or illness grades relies on cell-level data. So it is important to analyze individual cells for precise diagnosis. We investigated the performance of holistic features, topological features and combination of both in lung cancer pathology and proposed a quantitative image analysis framework [13]. As not may works have been done on classification and subtype recognition of NSCLC (Non-small Cell Lung Cancer), our work speaks for itself, and the main contribution being individual cell level analysis based on local and architectural features. The results from the experiments show that the architectural or topological features, plays a vital role in the differentiation of NSCLC as Adenocarcinoma and Squamous Cell Carcinoma. This work is done in hope for improving the efficiency of diagnosis of Non-small cell lung cancer. This would help in a quick detection, early diagnosis of NSCLC and its subtypes, thus aiding in the treatment of the patients. As we know early diagnosis is a key in the treatment of lung cancer, we aim to provide our support and aid in the process. Our proposed framework would benefit both the expert pathologists and the patients.

In spite of the fact that topological features give essential information for the purpose of classification [13], they exceptionally rely on upon the detection precision of nuclei centroids.In our future work, we plan to examine exhibitions when nuclei centroids are given via computerized systems. Additionally, motivated by recent structure sparse learning systems [34, 35, 36, 37, 38, 39, 40], low rank modeling tech-

nique [41, 42] and preconditioning efficient methods [43, 44, 45, 46], we can decrease the measurements of elements i.e. dimensions of features and further accelerate the whole process and enhance precision by using such feature selection methods.

REFERENCES

[1] H. Wang, F. Xing, H. Su, A. Stromberg, and L. Yang, "Novel image markers for non-small cell lung cancer classification and survival prediction," *BMC Bioinformatics*, vol. 15, no. 1, p. 310, 2014. [Online]. Available: http://www.biomedcentral.com/1471-2105/15/310

[2] V. K. Anagnostou, A. T. Dimou, T. Botsis, E. J. Killiam, M. D. Gustavson, R. J. Homer, D. Boffa, V. Zolota, D. Dougenis, L. Tanoue, *et al.*, "Molecular classification of nonsmall cell lung cancer using a 4-protein quantitative assay," *Cancer*, vol. 118, no. 6, pp. 1607–1618, 2012.

[3] Adenocarcinoma. [Online]. Available: http://www.cancercenter.com/terms/adenocarcinoma/

[4] Squamous cell carcinoma. [Online]. Available: http://lungcancer.about.com/od/typesoflungcancer/a/Squamous-Cell-Carcinoma-Of-The-Lungs.htm

[5] Y. Li, C. Chen, F. Yang, and J. Huang, "Deep sparse representation for robust image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4894–4901.

[6] Y. Li, C. Chen, W. Liu, and J. Huang, "Subselective quantization for large-scale image search," *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[7] Y. Li, F. Nie, H. Huang, and J. Huang, "Large-scale multi-view spectral clustering via bipartite graph," *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[8] Y. Li, C. Chen, and J. Huang, "Transformation-invariant collaborative sub-representation," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 3738–3743.

[9] Y. Li, W. Liu, and J. Huang, "Scalable sequential spectral clustering," *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[10] X. Zhang, H. Su, L. Yang, and S. Zhang, "Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5361–5368.

[11] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, no. 6, pp. 610–621, 1973.

[12] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot, and A. Madabhushi, "Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 642–653, 2010.

[13] J. Yao, D. Ganti, X. Luo, G. Xiao, Y. Xie, S. Yan, and J. Huang, "Computer-assisted diagnosis of lung cancer using quantitative topology features," in *Machine Learning in Medical Imaging*, ser. Lecture Notes in Computer Science, L. Zhou, L. Wang, Q. Wang, and Y. Shi, Eds. Springer International Publishing, 2015, vol. 9352, pp. 288–295. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24888-2_35

[14] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[15] S. Wienert, D. Heim, K. Saeger, A. Stenzinger, M. Beil, P. Hufnagl, M. Dietel, C. Denkert, and F. Klauschen, "Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach," *Scientific reports*, vol. 2, 2012.

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

[17] L. Grady and E. L. Schwartz, "Isoperimetric graph partitioning for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 3, pp. 469–475, 2006.

[18] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, *et al.*, "Cellprofiler: image analysis software for identifying and quantifying cell phenotypes," *Genome biology*, vol. 7, no. 10, p. R100, 2006.

[19] Image segmentation. [Online]. Available: https://en.wikipedia.org/wiki/Image_segmentation

[20] Voronoi diagram. [Online]. Available: https://en.wikipedia.org/wiki/Voronoi_diagram

[21] Minimum spanning tree. [Online]. Available: https://en.wikipedia.org/wiki/Minimum_spanning_tree

[22] Nlst. [Online]. Available: https://en.wikipedia.org/wiki/National_Lung_Screening_Trial

[23] Sensitivity and specificity. [Online]. Available: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[24] Cross validation. [Online]. Available: https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[25] Knn. [Online]. Available: http://chem-eng.utoronto.ca/ datamining/Presentations/KNN.pdf

[26] Knn. [Online]. Available: https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/

[27] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi, "Multifeature prostate cancer diagnosis and gleason grading of histological images," *Medical Imaging, IEEE Transactions on*, vol. 26, no. 10, pp. 1366–1378, 2007.

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[29] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *IEEE International Symposium on Biomedical Imaging*. IEEE, 2008, pp. 496–499.

[30] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[31] Naive bayes. [Online]. Available: http://software.ucv.ro/ cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf

[32] Z. Xu and J. Huang, "Efficient lung cancer cell detection with deep convolution neural network," in *Patch-based Techniques in Medical Imaging*, ser. Lecture Notes in Computer Science, G. Wu, P. Coup, Y. Zhan, B. Munsell, and D. Rueckert, Eds. Springer International Publishing, 2015, vol. 9467.

[33] H. Pan, Z. Xu, and J. Huang, "An effective approach for robust lung cancer cell detection," in *Patch-based Techniques in Medical Imaging*, ser. Lecture Notes in

Computer Science, G. Wu, P. Coup, Y. Zhan, B. Munsell, and D. Rueckert, Eds. Springer International Publishing, 2015, vol. 9467.

[34] J. Huang, "Structured sparsity: Theorems, algorithms and applications," Ph.D. dissertation, New Brunswick, NJ, USA, 2011.

[35] C. Chen, Y. Li, and J. Huang, "Calibrationless parallel MRI with joint total variation regularization," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. Springer, 2013, pp. 106–114.

[36] J. Huang, S. Zhang, and D. Metaxas, "Efficient mr image reconstruction for compressed mr imaging." *Medical image analysis*, vol. 15, no. 5, pp. 670–9, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841511000843

[37] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *The Journal of Machine Learning Research*, vol. 12, pp. 3371–3412, 2011.

[38] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2502–2514, 2015.

[39] C. Chen, Y. Li, and J. Huang, "Forest sparsity for multi-channel compressive sensing," *Signal Processing, IEEE Transactions on*, vol. 62, no. 11, pp. 2803–2813, 2014.

[40] C. Chen, Y. Li, W. Liu, and J. Huang, "Image fusion with local spectral consistency and dynamic gradient sparsity," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2760–2765.

[41] J. Yao, X. Liu, and C. Qi, "Foreground detection using low rank and structured sparsity," in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1–6.

[42] J. Yao, Z. Xu, X. Huang, and J. Huang, "Accelerated dynamic MRI reconstruction with total variation and nuclear norm regularization," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, vol. 9350, pp. 635–642.

[43] Z. Xu, Y. Li, L. Axel, and J. Huang, "Efficient preconditioning in joint total variation regularized parallel MRI reconstruction," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, vol. 9350, pp. 563–570.

[44] R. Li, Y. Li, R. Fang, S. Zhang, H. Pan, and J. Huang, "Fast preconditioning for accelerated multi-contrast MRI reconstruction," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Springer International Publishing, 2015, vol. 9350, pp. 700–707.

[45] C. Chen, Y. Li, W. Liu, and J. Huang, "SIRF: Simultaneous image registration and fusion in a unified framework," *arXiv preprint arXiv:1411.5065*, 2014.

[46] C. Chen, Y. Li, L. Axel, and J. Huang, "Real time dynamic MRI with dynamic total variation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 138–145.

BIOGRAPHICAL STATEMENT

Dheeraj Ganti joined the University of Texas at Arlington in Spring 2014. He received his B.Tech in Information Technology from GITAM University Visakhapatnam in 2013.In USA, he worked as a graduate research assistant at UTA as System administrator in the department of Computer Science. His current research interests include computer vision, image processing, big data, data mining and analytics.