Fall 2014

Uterine cervical cancer histology image feature extraction and classification

Koyel Banerjee

UTERINE CERVICAL CANCER HISTOLOGY IMAGE FEATURE EXTRACTION AND CLASSIFICATION

by

KOYEL BANERJEE

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

2014

Approved by

R. Joe Stanley, Advisor

Randy H. Moss

William V. Stoecker

ABSTRACT

The current study presents the investigation and development of image processing, computational intelligence, fuzzy logic, and statistical techniques for different types of data fusion for a varied range of applications. Raw data, decision level and feature level fusion techniques are explored for detection of pre Cervical cancer (CIN) grades from digital histology images of the cervical epithelium tissues.

In previous research, an automated, localized, fusion-based approach was investigated for classifying squamous epithelium into Normal, CIN1, CIN2, and CIN3 grades of cervical intraepithelial neoplasia (CIN) based on image analysis. The approach included medial axis determination, vertical segment partitioning as medial axis orthogonal cuts, individual vertical segment feature extraction and classification, and image-based classification using a voting scheme to fuse the vertical segment CIN grades. This paper presents advances in medial axis determination, epithelium atypical cell concentration feature development and a particle swarm optimization neural network and receiver operating characteristic curve technique for individual vertical segment-based classification. Combining individual vertical segment classification confidence values using a weighted sum fusion approach for image-based classification, exact grade labeling accuracy was as high as 90% for a 62-image data set.

ACKNOWLEDGMENTS

I would like to acknowledge, foremost, my advisor Dr. R. Joe Stanley for providing me the opportunity to pursue graduate studies. I am thankful for his constant support– both professional and personal to follow my interests. Dr. Stanley's teaching of several innovative ways of problem solving and out-of-the-box thinking has helped fuel my research to various interesting and novel directions. I thank Dr. Soumya De, whose continued guidance and initial work on this project established the platform from where I could carry on. I also wish to thank my committee members Dr. Randy H. Moss and Dr. William V. Stoecker for their guidance and support.

I would like to thank my parents for dedicating their lives to make their children happy and my brother for being my friend and a life-long source of guidance and inspiration. I would like to thank all my friends and colleagues Cheng Lu, Peng Guo and Xiao Pan for their company, help and cooperation. I would like to thank my friends Tamal, Doyal and Kancy for giving me a chance to share my good and hard times in Graduate School.

This material is based on work supported by the National Library of Medicine (NLM).

TABLE OF CONTENTS

Page

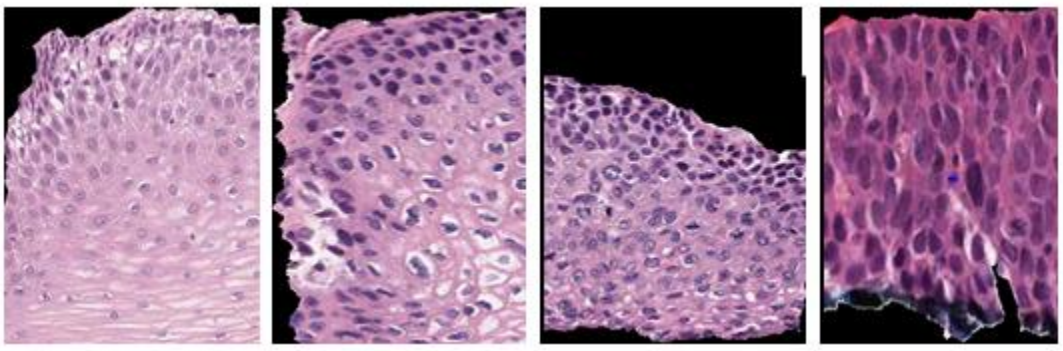
LIST OF ILLUSTRATIONS

LIST OF TABLES

Page

1. INTRODUCTION

Annually, there are 400,000 new cases of invasive cervical cancer out of which 15,000 occur in the U.S. alone [1]. Diagnosis for cervix tissue abnormalities is commonly based on performed procedures, including Pap test, a colposcopy to visually inspect the cervix, and visual inspection of histology slides when biopsied cervix tissue is available. Expert pathologist visual inspection of histology slides has been used as a standard of diagnosis [2]. Pathologists commonly assess Cervical Intraepithelial Neoplasia (CIN), provide diagnoses related to CIN and its various grades based on the visual interpretation of histology slides [3–7]. As part of the pathologist diagnostic process, Cervical intraepithelial neoplasia (CIN) is a pre-malignant condition for cervical cancer in which the atypical cells are examined in the epithelium [3] and is commonly assessed in the visual inspection of histology slides [3,7]. CIN is categorized into grades including Normal, CIN1 (mild dysplasia), CIN2 (moderate dysplasia), CIN3 (severe dysplasia) [3–5]. Fig.1.1 presents image examples of the different CIN grades, as determined by an expert pathologist.



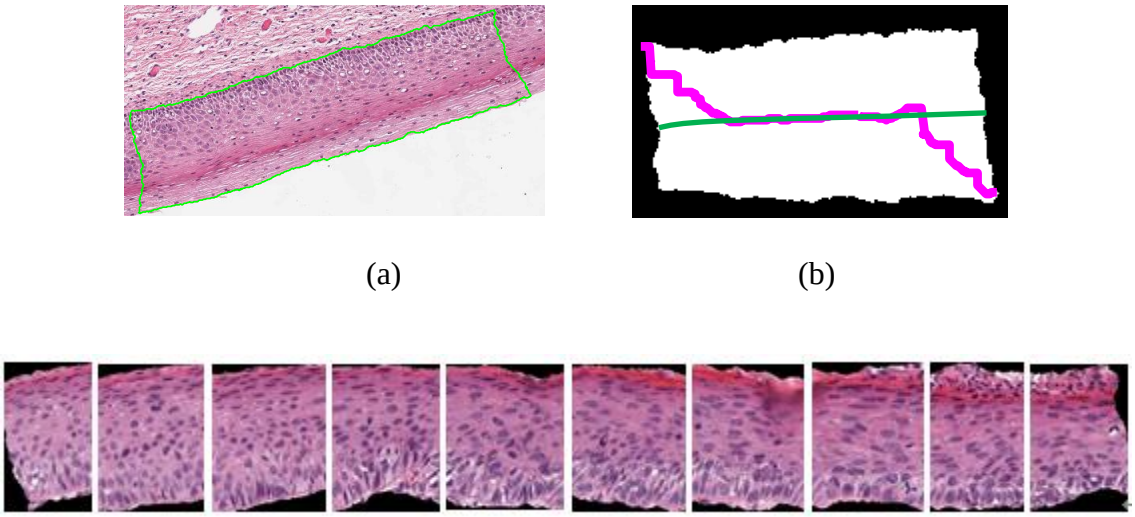
(a) (b) (c) (d)

Figure 1. 1. CIN grading label examples. (a) Normal, (b) CIN 1, (c) CIN 2, (d) CIN 3.

In previous research, a localized, fusion-based approach was investigated for cervix histology image analysis to classify the epithelium region into different CIN grades, as determined by an expert pathologist [8]. There have been several studies using digital pathology to complement traditional pathologists' decisions for CIN assessment [9-16]. Our approach built off of Keenan's automated squamous epithelium region-based image system for CIN classification based on nuclear center determination for use with Delaunay triangulation [8, 11]. Numerous image analysis studies have explored for CIN grade classification, including: 1) nuclei determination with Delaunay triangulation for feature analysis and CIN classification [10,14], 2) nuclei characterization using texture features [12,13], 3) nuclei morphological feature development [10,12,13,16,17,18], 4) histological feature development [15], 5) whole cervix slide histology image analysis with epithelium segmentation, medial axis determination and block wise morphological feature analysis [17,18], 6) heterogeneous epithelium region CIN grade analysis 7) nuclei morphological feature development [for nuclei char, including: a) dividing the image into horizontal compartments and computing morphological features from intra-compartment triangles [10], b) nuclei determination using Watershed segmentation [14], c) for image analysis techniques for computer-assisted CIN classification, including nuclei determination for Delaunay triangulation. Previously, computer-aided methods (digital pathology) have been investigated to augment 85 pathologists' decisions (traditional pathology) for CIN diagnosis [10–16]. Our study depends on morphological feature extractions to be used as an input to our PSO based Neural-net classification as followed from Guillaud, who extracted texture features from the epithelium region to estimate the absolute intensity and density levels of the nucleus [13]. Morphological

features were also extracted to estimate the nuclear shape, size and boundary irregularities [14]. Miranda et al. [15] determined the nuclei in the epithelium using a Watershed segmentation method followed by Delaunay Triangulation to facilitate CIN analysis, the concept of which was found to be very useful in our case study. This method was known for uniquely assigning CIN grade labels based on triangles using the Delaunay Triangulation method, instead of making a CIN grade decision on the whole epithelium image.

In previous research, histology image analysis of the epithelium for CIN classification was performed using the following steps: 1) medial axis determination based on the distance transform approach, 2) partitioning the epithelium into ten vertical segment as medial axis orthogonal cuts, 3) extracting features from the individual vertical segments, 4) classifying individual segments using Support Vector Machine and Linear Discriminant Analysis classifiers. The epithelium image analysis approach Includes medial axis determination, vertical segment partitioning as medial axis orthogonal cuts, individual vertical segment feature extraction and classification, and image-based classification using a voting scheme fusing the vertical segment CIN grades. CIN grade classification of each block was done by an SVM-based classifier using the features extracted from the blocks [17-19]. 5) Wang et al. [5] and Marel et al. [20] paved the process of development of a localized, fusion-based approach to provide the capability to address feature and diagnostic variations in different portions of the epithelium and combine those variations into a single diagnostic assessment. Image-based classification using a voting scheme fusing the vertical segment CIN grades [8, 21, 22, 23]. Figure 1.2 shows the overview for segmenting the epithelium region into 10 segments.



(a) (b)

(c)

Figure 1. 2. Overview of process (a) Original image, (b) medial axis image based on distance transform approach, (c) partitioning of epithelium into 10 vertical segments.

As previously documented, the large number of nuclei and other morphological parameters such as cytoplasm homogeneity, thickness of the nuclear membrane, and the presence of nucleoli contribute to the complexity and difficulty of visually assessing the degree of CIN abnormality. This in turn may contribute to 82 diagnostic grading reproducibility issues and inter- and 83 intra-pathologist variation [6–8].

In this research, using the feedback from two expert pathologists for CIN evaluation of histology images, feature development has focused on characterizing nuclear atypical, which is associated with nuclear enlargement related to differences in shapes, sizes and distribution of the nuclei present within the epithelium. The goal for this research is to extend previous work [8] in the development of new algorithms for medial axis determination, feature extraction, and increase the accuracy of the image classification by using different approaches in every step of digital image classification processes. The pre and post image processing techniques were built on [24], A new set of

features involving the nuclei(atypical cell) density ratio and light-area content of the images are being studied, apart from the previously used common nuclear-cytoplasmic ratio feature [19,20,25,26,27] as shown in Figure 1.3, for CIN classification.

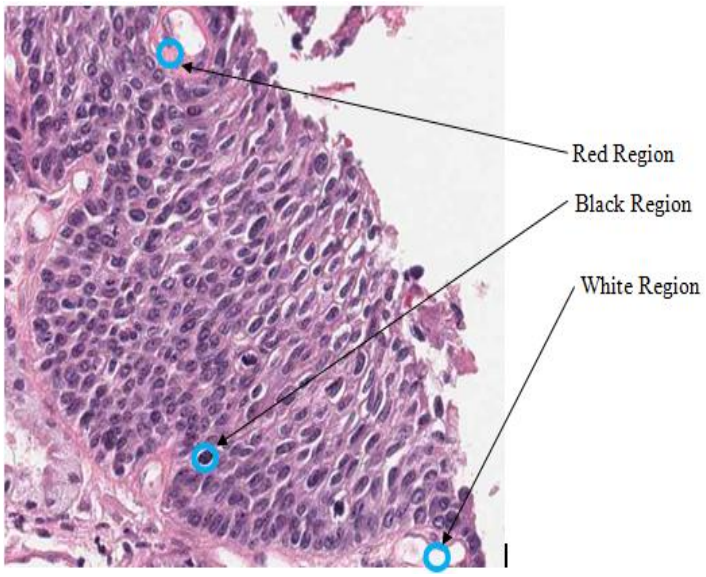


Figure 1. 3. Representative of color regions with previously used features..

The premise for this feature investigation is that there is an increase in the number of nuclei and a corresponding decrease in the light-area contents in the progression from Normal to CIN3. Therefore, normal and CIN1 images typically have a lower concentration of nuclei compared to CIN2 and CIN3 grades, which is illustrated in Figure 1.4.

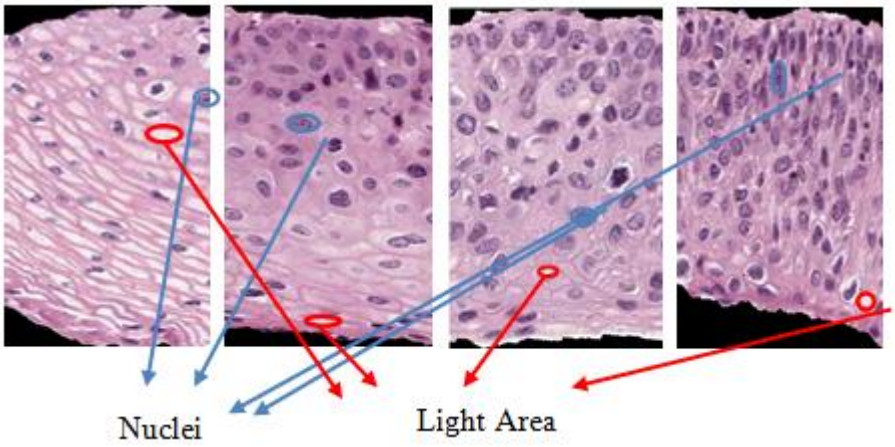
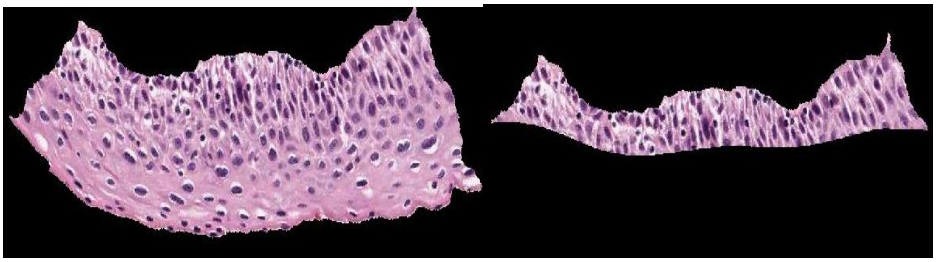


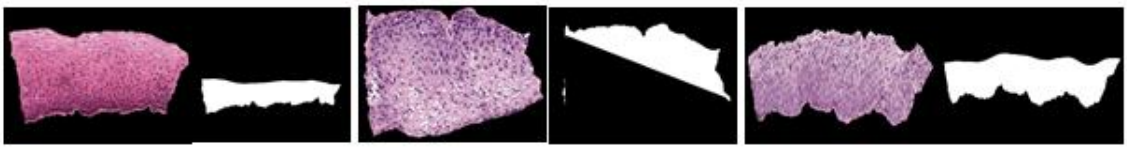
Figure 1. 4. Normal, CIN I, CIN II, CIN III stage images (from left to right in order).

Among the other features developed were Layer-by-Layer triangle features, which use the centers of detected nuclei for Delaunay triangle formation. One of the recently used methods segmented the nuclei from the epithelium using a K-means clustering and graph-cut segmentation method. The approach used a decision tree for CIN grade classification with empirically determined rules [16]. The growth of the abnormal nuclei starts either from the top or the bottom end of the epithelium lining. This is referred to as a basal membrane, as shown in Figure 1.5. It has been observed that the basal membrane tends to differ in width with different CIN grades. It was investigated that the more advanced the grade of CIN is, wider the Basal membrane tends to become also the density of the atypical cells increases with sharp elongation evident in their morphology. In other words, the basal membrane marks where the nuclei start to grow across (top/bottom) the whole epithelium. Figure 1.6 shows how the basal membrane width varies with respect to various CIN grades.



(a) (b)

Figure 1. 5. Basal membrane example. (a) Original cropped epithelium. (b)Basal membrane part from (a).



(a) (b) (c)

Figure 1. 6. Basal membrane masks growth with different grades of CIN. (a) Normal and Normal basal mask. (b) CIN 1 and its basal mask. (c) CIN 3 and its basal mask.

This research extends algorithms, techniques, and algorithms to estimate the epithelium medial axis, partition the epithelium into 10/5 vertical segments, compute features for each individual segment, classify each segment, fuse the 10/5 vertical segment designations (normal, CIN1, CIN2, CIN3) to a single classification for the entire epithelium (image-based classification). In this thesis, new approaches are presented for: 1) medial axis determination, 2) features computed from the individual segments partitioned within the epithelium region to characterize the basal membrane, nuclei distribution and light areas, 3) an evolutionary algorithm approach for feature selection from the individual segments, 4) a particle swarm optimization neural network technique for individual segment classification, and 5) a weighted sum image-based (epithelium)

CIN classification algorithm for combining the individual segment neural network classifier outputs research for the different steps of the epithelium analysis process.

The remainder of this thesis is organized as follows. Section 2 presents the methodology. Section 3 presents the medial axis detection and vertical segment determination algorithms. Section 4 presents the vertical segment features developed, with emphasis on the nuclei and light area detection and segmentation. Section 5 gives explanations of Nuclei and light area extraction process. Section 6 presents the Basal membrane detection and feature sets. Section 7 presents all the feature groups together with a general discussion. Section 8 presents the neural network approach used for individual vertical segment classification in this research, the weighted sum fusion technique for image-based epithelium CIN grading, Section 9 describes the experimental results yielded. Section 10 presents the conclusions.

2. METHODOLOGY

The goal of this research is to classify the squamous epithelium regions from cervix histology images into different grades of CIN. In the research, 62 cervix histology images were obtained in collaboration with the National Library of Medicine (NLM), with the epithelium manually segmented and CIN grade classifications determined by an expert pathologist. The research presented in this thesis extends the study in [8] for the development of new image analysis and classification of individual vertical segments for whole image for CIN grade determination. Figure 2.1 shows the flowchart of the overall method developed in this study for CIN grade classification. This thesis also presents CIN grade comparative classification analysis for two expert pathologist CIN grading of the 62 image data set.

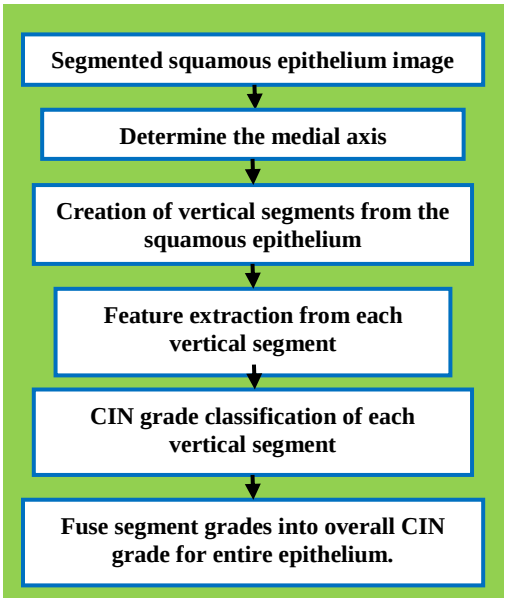


Figure 2. 1. Overview of CIN grade classification method developed in this study.

Figure 2.1 can be summarized in the following steps:

- Medial axis detection, locate the medial axis of the segmented epithelium region
- Image segmentation; divide the segmented image into 5 or 10 different vertical blocks along the medial axis.
- Feature extraction, extract features from each of the blocks in which is done by me with creating and testing several feature groups that help a lot in classification.
- Image-based classification, classify each of these segmented blocks into the different CIN grades.

3. MEDIAL AXIS DETECTION

The detailed medial axis determination for epithelium analysis for CIN grade classification is presented here. This approach for medial axis detection is an extension and improvement on the work done by De et al.[8]. This approach uses distance transform to estimate the interior 60% of the medial axis with the help of the surrounding bounding box and project the distance transform-based medial axis to the median bounding box points for the left- and right-hand end points (remaining left-hand 20% and right-hand 20% portions of the axis) [8,24]. However, the algorithm had difficulties finding the left- and right-hand portions of the axis in histology images with a somewhat rectangular epithelium region. Figure 3.1 shows two examples of improper medial axis estimation using the distance transform-based approach. The line shown in pink color is the detected medial axis using the distance transform approach while the line shown in green is the manually marked medial axis, which is the desirable medial axis. In order to address these limitations with medial axis determination, a bounding box-based method was explored.

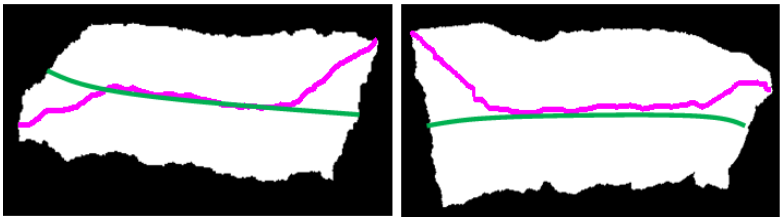


Figure 3. 1. Examples of incorrect medial axis estimation.

A bounding box-based medial axis estimation algorithm was investigated, as part of this research, based on a two-step method. While this method follows the original

distance transform fundamentals [8] to find out the centroid points at the end segments for correction to prevent bending the medial axis along the edges, it also uses a second method based on the morphology of the epithelium area before making the decision about which method to follow for the axis detection.

The bounding box-based method uses ratio comparison of the number of nuclei distributed over 8 masks that are created from the bounding box and control points marked on it. Also for precision a 16 mask approach along with symmetry factor of the image was taken into consideration. The following Figure 3.2 explains the concept.

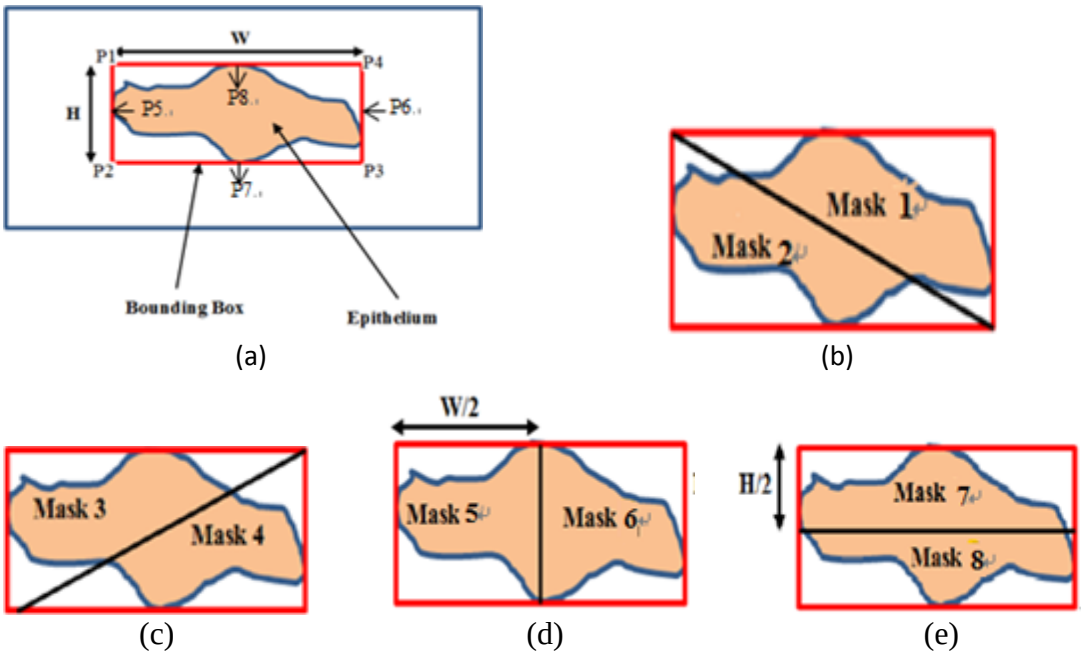


Figure 3. 2. Bounding box medial detection algorithm example (a) the control point on the bounding box. (b) Mask 1 & Mask 2 (c) Mask 3 & Mask 4 (d) Mask 5 & Mask 6

The masks are used for computing the ratios of the number of detected nuclei to the ratio of the area of the masks. The masks are namely obtained after dividing the epithelium area into 2 vertical segments, 2 horizontal segments and 2 diagonal segments

as shown in the Figure 3.2. The Control points used for determining the masks are shown as P1 through P6 as shown in Figure 3.2 (a). For example if the number of nuclei in mask 1 in Figure 3.2 is supposed to be n_1 and in mask 2 is n_2 then compute a ratio as f, which is a ratio number of nuclei in corresponding masks in an image. As seen from figure 3.2 there are 8 masks so 8 such ratios can be calculated starting from $f_1 \ldots f_8$, where, $f_1 = \frac{n_1}{n_2}$, f_2. Other ratios for normalizing the n (n_1, n_2..) values were also computed like $n_1 = \frac{n_1}{\max(n)}$ where n is an array containing ($n_1, n_2 \ldots n_8$) where each n1 through n8 represent the total number of nuclei detected in the corresponding masks.

Finally this result was multiplied by the eccentricity value of the particular mask. The equation guiding the detection of the medial axis can be given as $v_1 = \frac{n_1 f_1}{e_1}$, $v_2 = \frac{n_2 f_2}{e_2} \ldots v_3 = \frac{n_3 f_3}{e_3}$ where n symbolizes the array containing the normalized ratios of $n_1, n_2, n_3 \ldots . n_8$, f symbolizes the array containing $f_1 f_2 \ldots f_8$ and $e_1 = \frac{e_1}{e}$ contains the normalized eccentricity measures of the 1st of the eight masks and e represents the eccentricity value of the entire tissue slide. The idea is that whichever position gives the maximum value by this computation is the medial axis position. Some of the medial axis image examples are stated below.

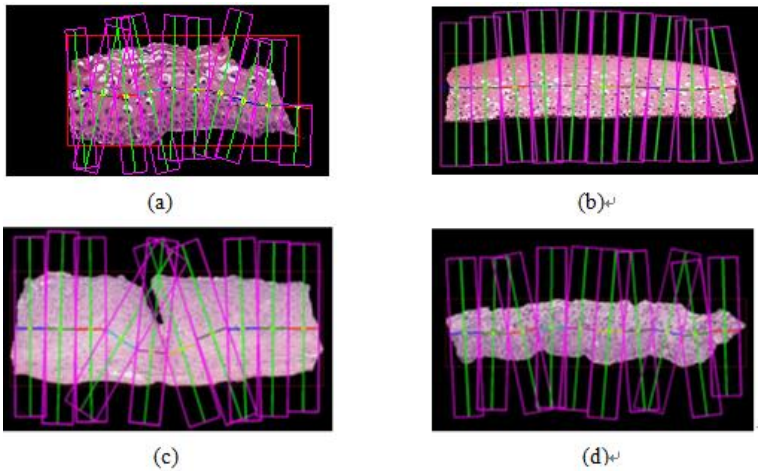


Figure 3. 3. Example of medial axis found using bounding box-based algorithm. (a) & (b) show bounding box method with axis extending beyond epithelium region. (c) & (d) show bounding box updated algorithm with medial axis contained with epithelium region.

As seen from the right-hand medial axis example image in Figure 3.3, some of the segments extend way out of the actual epithilial region. This is because in this approach points on the bounding box which is exterior to the epithilium region was used. The classification process discards such segments and take up only valid segments out of the ten segments for an image for decision of CIN label which sometimes degraded the efficiency of the grade estimation. To overcome this problem the algorithm was again modified so that the medial axis only extends for the part in the bounding box where an epithelium is present.

4. IMAGE SEGMENTATION

Vertical image segmentation is considered a critical step for this research because most of the images contain a large number of data points, which gives a lot of burden on the future step, feature extraction. Also, by visually examining the CIN grades segment wise, different portion of the same image might be considered cancerous. Among normal, CIN 1, CIN 2, and CIN 3 grades, one single image might contain multiple grades, which gives inconsistent results for the classification process. Therefore, segmenting the images into small segments, more testing images will be collected for further research and each segmented image gives a fairly consistent CIN grade. This is further discussed by Lu [21]. The initial manual segmentation of the epithelium from the NLM digital histology images were done manually using .Xml files provided by the pathologists.

4.1 VERTICAL IMAGE SEGMENTATION

Based on the approach presented by Lu [21], Figure 4.1 shows bounding box regions based on breaking the medial axis within the epithelial region into 10 approximately equidistant segments. The algorithm for extracting the 10 vertical segments is presented in detail in [8,21,22]. A summary of the method is presented here. The orientation of each segment is determined by taking all of the medial axis points and estimating the slope of the medial axis within each of these ten segments along the medial axis. Points within each of the segments are curve-fitted using a least-squares approach [28] to fit the points along the medial axis orientation is determined so that a bounding box can be generated. For each epithelial region, the 10 bounding box areas (regions) are extracted to be used for feature and classification analysis. In the current

research 5 and 10 vertical segments were examined for CIN discrimination analysis. Note that the algorithm to partition the epithelium region into 5 vertical segments is similar to the 10 vertical segment algorithm presented. The various image processing techniques used for the following feature extraction process was built on [29]. The various steps of the method to extract 10 vertical segments method are shown in Figure 4.2 below.

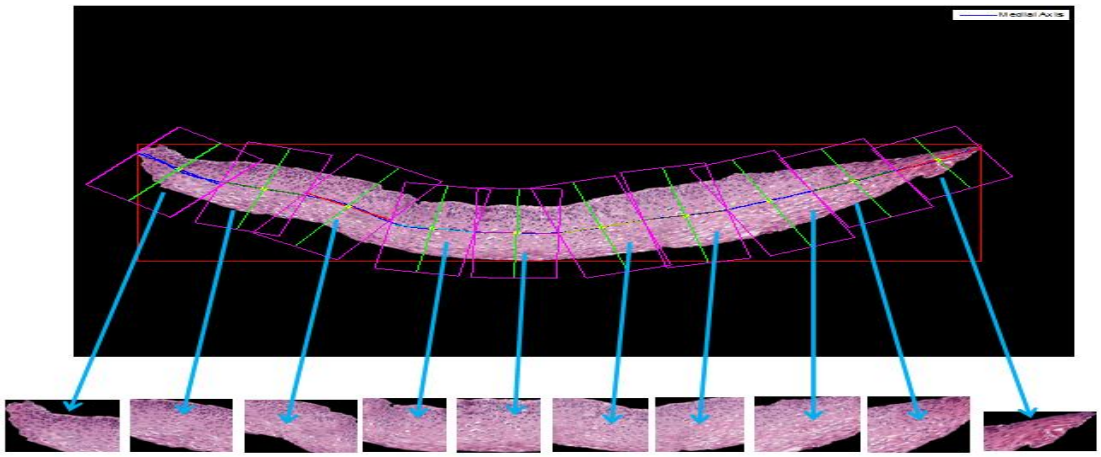
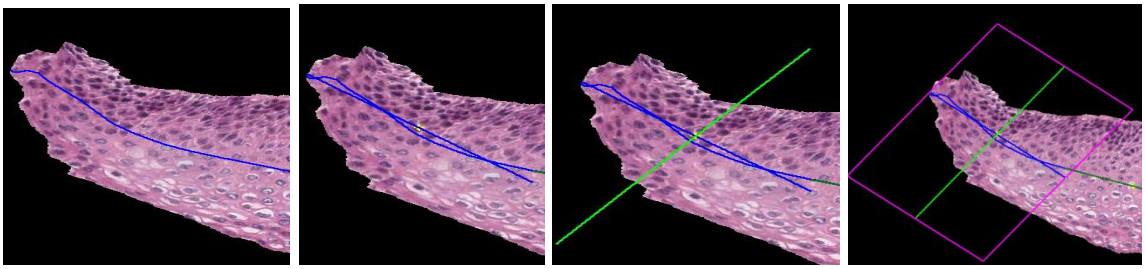


Figure 4. 1. Example of medial axis broken into 10 segments with bounding boxes shown/determined for each segment.



(a)	(b)	(c)	(d)
The portion of the medial axis within the first segment	The least squares fit line obtained from the medial axis points.	The perpendicular of the least squares fit line obtained to generate the bounding box	The bounding box is generated using the information obtained from Steps (a)-(c).

Figure 4. 2. The various steps in creating the ten different segments from the epithelium region.

5. NUCLEI AND LIGHT AREA DETECTION

The nuclei and light area detection and feature extension was crucial for this study. Certain other features called as Basal membrane features were also extracted and have been discussed in the following sections. The nuclei feature extraction was mainly developed by Lu [21] and Guo [23]. This automated algorithm extracts nuclei area and light area from the epithelium. Each given JPG image is extracted out from the data contains of an XML file, which marks the boundary position for the squamous epithelium of a greater tissue slide. The boundary position differentiates the useful area and outside area. As shown in Figure 5.1, the light area above the green box is not considered in the region of interest, since it is not a part of the squamous epithelium.

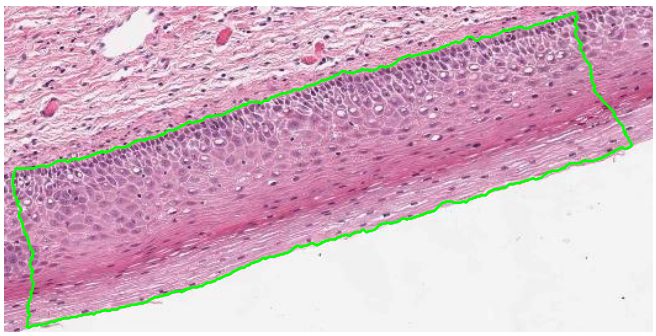


Figure 5. 1. Original large image with green boundary.

Previous research has explored dividing the ROI into 10 uniform width vertical segments (sub-images) along the medial axis determined with the help of the bounding box [8]. In this study, epithelium region analysis and classification was investigated for 10 and 5 vertical segment cases. In continuation, a data set of 620 sub-images (10 vertical

segments from each of the 62 images) from the original 62 images is created which were obtained from the National Library of Medicine (NLM). The Figure 5.2 shows some samples of sub-images.

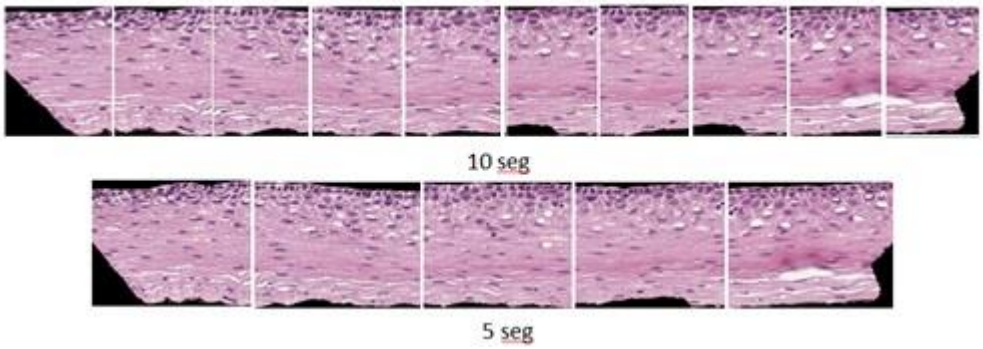


Figure 5. 2. Dividing the original image into sub-images of 10 and 5.

5.1. NUCLEI PRE-PROCESSING AND SEGMENTATION

This study investigates a new nuclei detection algorithm based on epithelium image pre-processing to make the image enhanced for nuclei detection. These pre-processing procedures mainly include averaging, image sharpening, histogram-equalization, high frequency boosting, etc.

In this segmentation process, image enhancement using High-boost Filtering is used to improve the contrast between the nuclei and the background for successful nuclei detection followed by histogram equalization to bind the pixel values in between 0 to 255. After testing a portion of the nuclei detection code which was supplied by NLM certain progresses such as clustering, holes filling, thresholding were studied. The following figure shows the process in Figure 5.3. To segment the nuclei from the tissue, K-means clusters are calculated from the given histology image after certain pre-

processing like High-Boost filtering and Histogram Equalization [21]. Since the contrast of the images is improved, the nuclei detection code can produce a better result. It was experimentally determined that processing the red layer of the RGB image accounted for the best result. After examination, the green and blue layers give similar but slight different resulting images as shown in Figure 5.4.

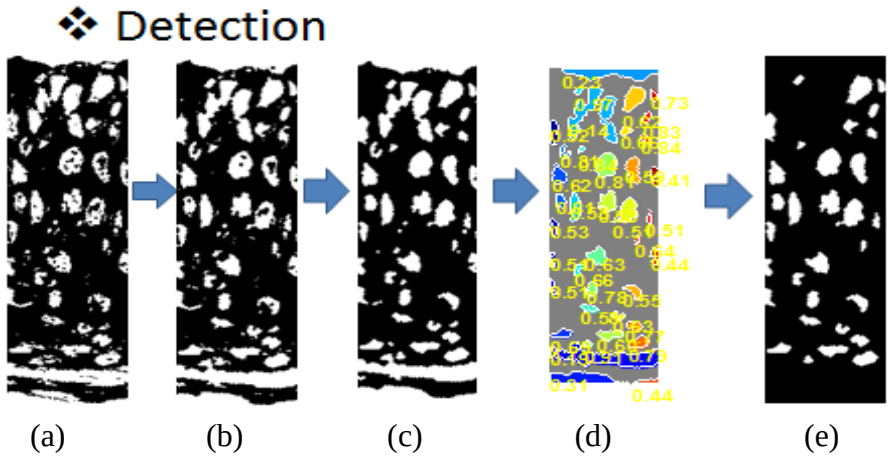


Figure 5. 3. Nuclei detection progress. (a) Original mask (b) & (c) Holes filling

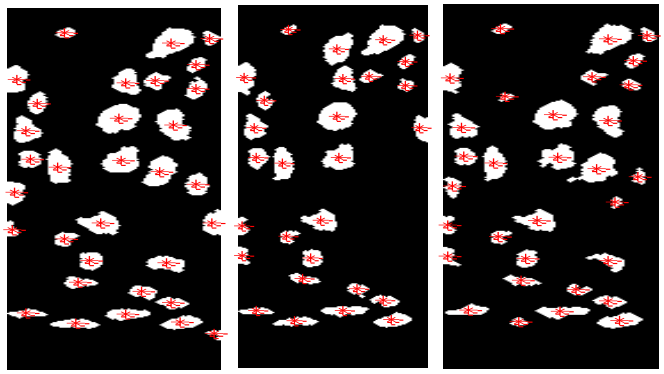


Figure 5. 4. Result images in three different color channels. (a) Red layer, (b) green layer, (c) blue layer.

5.2 LIGHT AREA SEGMENTATION

This involves extracting the lighter, clear areas found in the epithelium. The challenge that goes with extracting the light area regions from the original image is mainly the color and intensity variations. Often the light areas are mistaken for white areas which are not the case. The light areas may appear white to the human eye, but the light areas tend to be more on the tail on the histogram where there is the concentration of light areas or high intensity values. Figure 5.5 presents an example of the histogram of an epithelium region with several light areas.

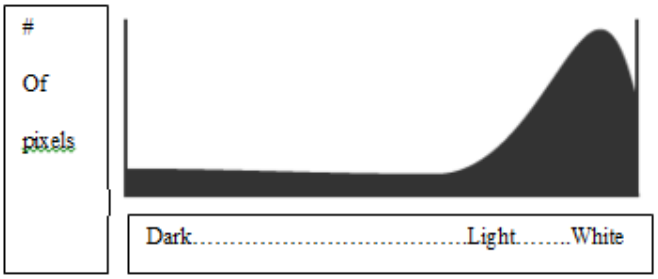


Figure 5. 5. Histogram of an image with several light areas.

In addition, the light areas do not have a pre-defined shape like the nuclei so it cannot be taken into account the shape/morphology of these regions. Therefore, to avoid such shortcomings an attempt to process these regions in the color plane was done taking into account the a-plane and b-plane and discarding the L-plane. The three coordinates of CIELAB represent the lightness of the color (L* = 0 yields black and L* = 100 indicates diffuse white; specular white may be higher), its position between red/magenta and green (a*, negative values indicate green while positive values indicate magenta) and its position between yellow and blue (b*, negative values indicate blue and positive values indicate yellow). The L-plane provided the best visual results of the 3 planes examined.

The following outlines the methods undertaken to segment the histology images:

1. The image was converted from RGB color space to L*a*b color space and taking out the luminance components that is the L plane for working further on it.

2. Contrast enhancement is performed using adaptive histogram equalization on the L plane image as an alternative to using 'histeq'. While 'histeq' works on the entire image, 'adapthisteq' operates on small regions in the image, called tiles, the default value for tiles is used here. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches a specified histogram. After performing the equalization, 'adapthisteq' combines neighboring tiles using bilinear interpolation to eliminate artificially induced boundaries. Adaptive histogram equalization yields better results than simple histogram equalization by the factor that the previous one equalizes the histogram with respect to 256 bins while the later only 64. The target histogram waveform was set to default which is uniform or in other words flat shaped. Contrast enhancement is performed on the image obtained from step 1 so that the light areas appear lighter and the dark areas appear darker for facilitating the extraction.

3. Thresholding is performed on the contrast enhanced image from step 2 with a value of 0.6 as obtained experimentally by analysis of the images from the data set. This step is used to eliminate the very dark nuclei regions, leaving behind the lighter nuclei and epithelium along with the light areas.

4. A modification of the Matlab K-means clustering algorithm is used to segment the light areas. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells.

A Voronoi diagram is equivalent to dividing the space into cells with each cell starting as a seed and expanding to a region. It is a dual of the Delaunay's triangulation procedure as found in Matlab. This is crucial for segmenting the light areas since these areas do not have either a fixed color value or a fixed contour or shape. The cluster image was graded into 4 observations depending on the median intensity values it represented as performed by a small 'comparison of value' routine. The cluster having the highest intensity value scores for the lighter most regions in the epithelium. The input to the K-means algorithm was the thresholded histogram equalized epithelium. The number of clusters was set to 4 as experimentally determined after certain trial and error analysis.

5. After the 4 clusters are obtained, it remains to be investigated which of the 4 clusters refers more closely to the light area regions. For this we do averaging of pixel values over each of the clusters. The one which has highest value should be consisting of the highest number of light areas and is of importance. After the cluster with the highest value is projected as a binary image it serves as the mask image for the light area regions. Afterwards, to get rid of the finer connected light regions, a morphological dilation is presented followed by erosion with "disk" as structure element and 2 as radius. Then, 'regionprops' is performed on the remaining image objects to keep only those light areas which are greater than an area of 100 pixels. 'regionprops' is a special command in Matlab which helps extract interesting region of interest properties which are eccentricity, centroid, area, bounding box etc.

The resulting large light areas are used for feature calculations for epithelium vertical and image-based classification which are shown in Figure 5.6.

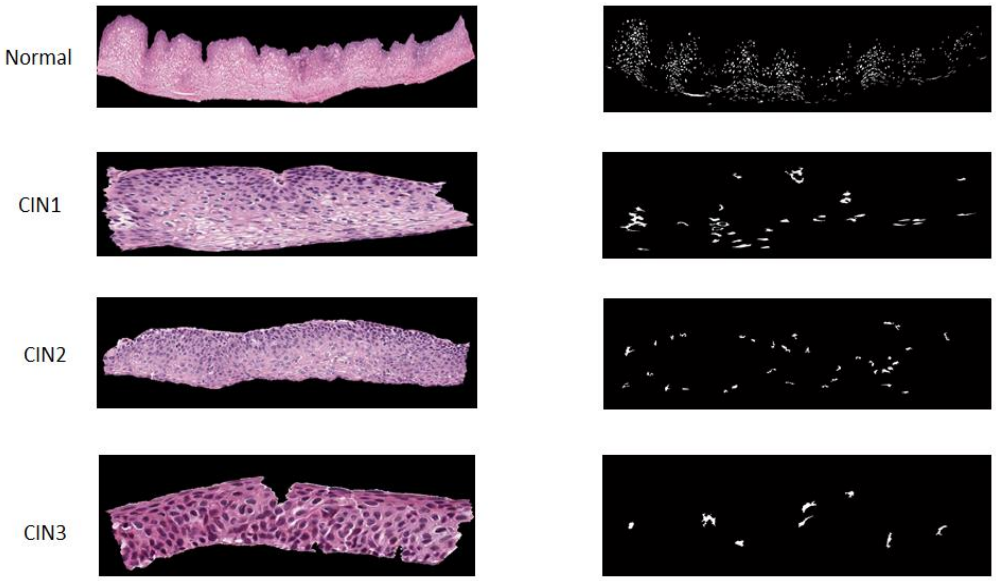


Figure 5. 6. Original epithelium and the light area segmentation.

6. BASAL MEMBRANE DETERMINATION AND SEGMENTATION

Depending on the epithelium orientation, the basal membrane is located at the bottom or top of the epithelium, typically containing numerous nuclei. Figure 6.1 shows an example of a basal membrane, where the basal mask is aligned towards the top side since the nuclei per square area ratio is highest in this entire region.

(a) Original image (b) Basal Mask (c) Basal Membrane

Figure 6. 1. generation of Basal mask.

In order to determine the width of the basal membrane the ratio of **nuclei per unit area** is estimated. The following approach was explored to determine the basal membrane location for feature extraction. This process is used only to determine the specific alignment of the membrane, if it is along the top or the bottom of the epithelium under consideration. If it is found to be along the top the image is flipped so that the alignment is along the bottom. The algorithm starts growing the membrane bottom up and therefore should start from $segment_1$. The entire epithelium region is divided into 10 horizontal segments. The rationale behind breaking up the entire epithelium into 10 horizontal segments and investigating the nuclei per unit area ratio is, if the entire region just above or below the medial axis is considered it can be subjected to a very high concentration of nuclei around the central part of the epithelium as compared to the edges giving rise to faulty results. Therefore experimentally investigating the epithelium in

horizontal strips proved more beneficial and accurate. Then calculate the difference of ratio for each strip is given as

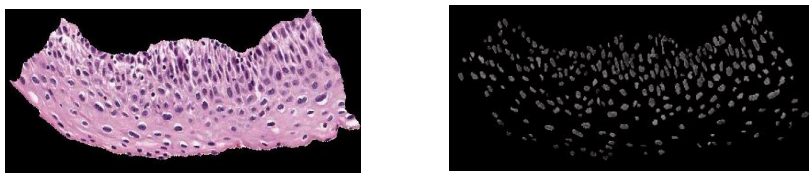
$diff1 = segment_1$ - $segment_{10}$ And see if diff1 is greater than zero. Where,

$segment_{i=} N_{n_i}/A_i$, *where, i=1,2,3…10*

$N_{n_i} = The\ total\ number\ of\ nuclei\ in\ each\ of\ the\ strips, where\ we\ divide\ the\ entire\ epithilium\ into\ 10\ horizontal\ strips\ with\ strip1\ being\ the\ bottommost\ and\ Strip\ 10\ being\ the\ topmost.$

$A_i = area\ of\ the\ entire\ epithilium$

If diff1 is a positive number, then given we take segment 1 to be the bottommost segment we know the image for Basal feature extraction is aligned the right way. If the opposite is true then the image is turned or flipped 180 degrees to have the basal membrane on the bottom side for extraction purposes. Based on the ratio diff1 the algorithm helps in locating the position of the membrane (towards the top or towards the bottom) and then dynamically detects the width of the basal membrane. Once the membrane has been extracted we perform a number of operations to extract certain features to help us determining the stage of CIN. Some of the features namely are: number of nuclei in the basal part/number of nuclei in the non-basal part of the epithelium, number of nuclei in the basal part/area of the entire epithelium, percentage of epithelium in the Basal Membrane. The following figures show the previous Figure 6.1 being broken into 10 strips with their corresponding nuclei mask. Details of the algorithm follow in the next sections.



(a) Main image and its corresponding nuclei mask

(b) Strip 1(the bottommost) and its corresponding nuclei mask

(c) Strip 2 and its nuclei mask

(d) Strip 3 and its nuclei mask

(e) Strip 4 and its nuclei mask

(f) Strip 5 and its nuclei mask

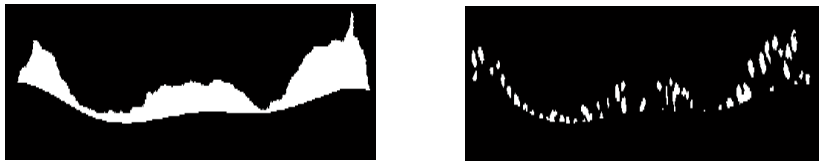
(g) Strip 6 and its nuclei mask

Figure 6. 2. Determination of the Basal membrane.

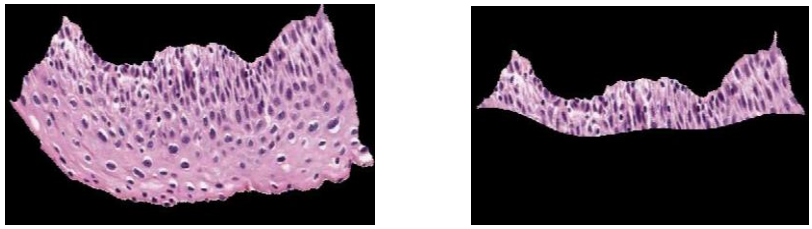
(h) Strip 7 and its nuclei mask

(i) Strip 8 and its nuclei mask

(j) Strip 9 and its nuclei mask



(k) Strip 10 and its nuclei mask



(i) Final computation of basal membrane

Figure 6.2 Determination of the Basal membrane (cont.)

From the above figures in Figure 6.2 it is clear that the nuclei concentration is higher towards the topmost part and hence the basal membrane extracted is from the top. Higher CIN grades (CIN2 and CIN3) typically have wider basal membranes than lower CIN grades (normal and CIN1). Therefore, this ratio/ feature provide a good estimation of the CIN grade. The following explains the algorithm in details.

1. The initial approach is to determine the location of the Basal membrane. The entire image was divided into 10 strips horizontally and stored separately where each strip represents 10% of the epithelium area.
2. For each strip identical set of operations were performed namely: creating the nuclei mask for that particular segment, finding out the number of nuclei, finding out the area of that entire strip, finding out the area within that strip populated by the nuclei.
3. Once we have these values a new ratio was created by dividing the number of nuclei belonging to a particular strip by the area of that entire strip. So if this ratio is named R, we have corresponding similar R ratios for all 10 horizontal segments namely R_1, $\mathrm{R}_2 \ldots \mathrm{R}_{10}$ with R_1 being the bottommost segment ratio and R10 the topmost. Therefore, $R_1 = \frac{nuclei\ in\ segment_1}{area\ of\ segement_1}$. Each segment consists of 10% of the total epithelium region.
4. So we have in total 10 ratios for comparison.
5. Next we calculate the 1st difference of these ratios, for example like difference between $R_1 - R_{10}$, $\mathrm{R}_2 - \mathrm{R}_9$ ...etc and denote it as '1stdiff'. Therefore, $\mathrm{1stdiff}_1 = \mathrm{R}_1 - \mathrm{R}_{10}$, $\mathrm{1stdiff}_2 = \mathrm{R}_2 - \mathrm{R}_9$
6. The purpose of step 5 is twofold. Firstly it is helpful to detect the orientation of the membrane using the $\mathrm{1stdiff}_1$ which is precisely $R_1 - R_{10}$. The rest of the calculated 1stdiff ratios help to dynamically determine the width of the membrane. In order to grow the membrane the 1st priority should be to detect if the membrane is aligned towards the upper side or towards the lower side.
7. If the 1stdiff calculation in step 5 ($R_1 - R_{10}$) is greater than zero, it means that the nuclei concentration in the 1st epithelium segment is more than the nuclei concentration in the

10^{th} topmost segment. Therefore roughly it can be concluded that the position of the basal membrane is along the lowermost part. Then to grow the membrane dynamically we must start comparing and looking through the other 1stdiff ratios from the lowermost part and proceed towards the uppermost part. Another ratio computation is simultaneously investigated to make a decision about the membrane thickness. Let the new ratio,

Dif_i = $segment_{i+1}$ - $segment_i$ for i =1 to 9 is computed as follows i.e. $Dif_{1=}\, segment_2 – segment_1$, $Dif_{2=}\, segment_3 - segment_2$...

8. Henceforth when the word 'Difference set' would be used it will precisely mean this ratio set named as Dif_i.
9. The next step is to dynamically detect the basal membrane. If the position of the Basal membrane is not along the bottom part of the epithelium it is done so by flipping the image 180 degrees. The process is two phased:

a) Basal membrane is oriented towards the bottom part of the image:

Then we started comparing ratios from 'Dif' bottom to up

If Dif_i > 0 AND Dif_{i+1}> 0 where i = 1 is bottom

Segment. That is the nuclei concentration shows a steady increase.

Also if:

Dif_{i+1}> Dif_i OR Dif_{i+1}< Dif_i

Which means both consecutive ratios that are ($Dif_1 = segment_2 – segment_1$) and ($Dif_2 = segment_3 - segment_2$) are positive. In this case even if either one of the be greater or smaller than the other it signifies a rise in nuclei count and so the next strip is added to the existing previous updated strip from start. For example if strip1was the start then now

the current temporary detected basal membrane would be $segment_1 + segment_2 \ldots$ Or vice versa $segment_{10} + segment_9$ in case of a membrane aligned towards top.

The other cases to consider while growing the membrane would be if

$Dif_i > 0$ AND $Dif_{i+1} < 0$

That means although the difference between ($Dif_{1=} segment_2 – segment_1$) is increasing there is a fall in nuclei count between ($Dif_{2=} segment_3 - segment_2$). Then a note of how much the nuclei/area ratio is decreasing between $segment_3 \; and \; segment_2$ can help. If the Dif_{i+1} is greater than -0.1 then we keep adding the strips. If it is less than -0.1 then basal membrane growth is stopped. The value of -0.1 has been determined experimentally.

The last special case to be considered would be if

$Dif_i < 0$ AND $Dif_{i+1} > 0$

It is simply a case where the area of the starting strip might be very thin so that not much nuclei are detected. In such a case we go on to add the current strip to the existing strip and form the temporary basal membrane.

b) Basal membrane is oriented towards the upper part of the image:

The same set of process and comparison is applied for images where the basal membranes are found aligned towards the uppermost portion. Except for the special case where the cut off value is set to -0.1 previously is replaced by -0.08 as determined experimentally from the training image set.

Once the entire iteration completes over the given range of 1 to 10 segments we get the final basal membrane.

7. FEATURE DEVELOPMENT

This research extends the feature development as undertaken in previous studies [8, 21, 22, 23]. Table 7.1 summarizes all the features computed from the individual vertical segments previous research and developed in this research and applied to vertical segment and image-based CIN discrimination. Features developed in previous research from Table 7.1 include: Texture Features [8], Color Features [8], Triangle Features [8], and Correlation-based Features (WDD) [8]. Features currently under development includes Nuclei Features [21], Light Area Features [21], and Combined Features [21] summarized as follows. After both the nuclei features and the light area features were extracted, some new features were generated by using some of the new attributes from nuclei and light area features. Such features include the ratio between the number of light areas and the number of nuclei, and the ratio between total light areas and total nuclei areas. The following few sections provide detailed description of the new features.

Table 7. 1. Feature table, (a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (m) Basal.

(a) TEXTURE FEATURES

Label	Measure	Description
F1	Contrast of segment	Intensity contrast between a pixel and its neighbor over the whole image.
F2	Energy of segment	Entropy (squared sum of pixel values in the segment)
F3	Correlation of segment	Measure of how correlated a pixel is to its neighbor over the whole image.
F4	Homogeneity of a segment	Closeness of the distribution of pixels in the segment to the segment diagonal.
F5-F6	Contrast of GLCM	Contrast of the GLCM matrix obtained from the segment.

Table 7.1. Feature table, (a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (m) Basal (cont.)

F7-F8	Correlation of GLCM	Closeness of the distribution of elements in the GLCM to the GLCM diagonal.
F9-F10	Energy of GLCM	Sum of squared elements in the GLCM.
F11	Correlation	Closeness of the distribution of elements in the GLCM to the GLCM diagonal.

(b) COLOR FEATURES

F12	Percentage Red	Percentage of region that has the reddish pixels.
F13	Percentage White	Percentage of region that has the whitish pixels.
F14	Percentage Black	Percentage of region that has the blackish pixels.

(c) TRIANGLE FEATURES

F15	Average area of triangles	Average area of the triangles formed by using Delaunay triangulation on the nuclei detected.
F16	Std deviation of area of the triangles	Standard deviation of the area of the triangles formed by using Delaunay triangulation on the nuclei detected.
F17	Average edge length	Mean of the length of the edges of the triangles formed.
F18	Std deviation of edge length	Standard deviation of the length of the edges of the triangles formed.

(d) CORRELATION-BASED FEATURES

F19~F66	Weighted density distribution	Correlation of texture profile of the segment and correlation function.

Table 7.1. Feature table,(a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (m) Basal (cont.)

(e) NUCLEI FEATURES

F67	Average nuclei area	The ratio of total no of nuclei over total area covered by the nuclei
F68	Ratio of background area over nucleus area	The ratio of total Nuclei area over total non-nuclei area

(f) LIGHT AREA FEATURES

F69	Ratio RGB	average intensity of RGB Light area mask image over non-light area epithelium
F70	Ratio R	average intensity of R-plane in luminance image of the Light area mask image over background(non-light area epithelium)
F71	Ratio G	average intensity of G-plane in luminance image of the Light area mask image over background
F72	Ratio B	average intensity of B-plane in luminance image of the Light area mask image over background
F73	Ratio LUM	average intensity of L-plane in luminance image of the Light area mask image over background
F74	Unit size of light area	The number of light area over total area
F75	Ratio of light area over background area	ratio of total light areas over total background (Light) area

Table 7.1. Feature table,(a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (k) Basal (cont.)

(g)COMBINED FEATURES

F76	Ratio of light area number over nuclei number	The ratio of light area number over nuclei number
F77	Ratio Light over Nuclei	The ratio of total light areas over total nuclei area

(h) LAYER-BY-LAYER TRIANGLE FEATURES

F78-F80	Average area of triangles In upper, mid and lower layer	This is the average area of the triangles formed by using Delaunay triangulation on the cells detected, from the upper layer to the lower layer.
F81-F83	Std deviation of area of the triangles in upper mid and lower layer	This is the standard deviation of the area of the triangles formed by using Delaunay triangulation on the cells detected, from the upper layer to the lower layer.
F84-F86	Average edge length of the triangles in upper mid and lower layer	This is the mean of the length of the edges of the triangles formed, from the upper layer to the lower layer.
F87-F89	Std deviation of edge length of triangles in upper mid and lower layer	Standard deviation of the length of the edges of the triangles formed, from the upper layer to the lower layer.
F90-F92	Number of triangles in three layers divided by square root of epithelium area	Counting the number of triangles in three different layers
F93-F95	Number of triangles	The ratio of number of triangles over the area of the three different layers and the total triangle number over the total area in the last feature
F96	Average area of triangles In upper, mid and lower layer	This is the average area of the triangles formed by using Delaunay triangulation on the cells detected, from the upper layer to the lower layer.
F97	Std deviation of area of the triangles in upper mid and lower layer	This is the standard deviation of the area of the triangles formed by using Delaunay triangulation on the cells detected, from the upper layer to the lower layer.

Table 7.1. Feature table,(a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (k) Basal (cont.)

F98	Average edge length of the triangles in upper mid and lower layer	This is the mean of the length of the edges of the triangles formed, from the upper layer to the lower layer.
F99	Std deviation of edge length of triangles in upper mid and lower layer	Standard deviation of the length of the edges of the triangles formed, from the upper layer to the lower layer.
F100	Number of triangles in three layers divided by square root of epithelium area	Counting the number of triangles in three different layers
F101	Number of triangles	The ratio of number of triangles over the area of the three different layers and the total triangle number over the total area in the last feature
F102-F104	Average area of triangles In upper, mid and lower layer	This is the average area of the triangles formed by using Delaunay triangulation on the cells detected, from the upper layer to the lower layer.
F105-F107	Std deviation of area of the triangles in upper mid and lower layer	This is the standard deviation of the area of the triangles formed by using Delaunay triangulation on the cells detected, from the upper layer to the lower layer.
F108-F110	Average edge length of the triangles in upper mid and lower layer	This is the mean of the length of the edges of the triangles formed, from the upper layer to the lower layer.
F111-F113	Std deviation of edge length of triangles in upper mid and lower layer	Standard deviation of the length of the edges of the triangles formed, from the upper layer to the lower layer.

Table 7.1. Feature table,(a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (k) Basal (cont.)

F114-F116	Number of triangles in three layers	Counting the number of triangles in three different layers
F117-F119	Number of triangles over area of the layer	The ratio of number of triangles over the area of the three different layers and the total triangle number over the total area in the last feature
F120	Total number of triangles over total area	The ratio of number of triangles over the area of the whole epithelium
F121	Background area over total triangle area	The ratio of background area over the area of all the triangles
F122	Average area of triangles	This is the average area of the triangles formed by using Delaunay triangulation on the cells detected.
F123	Std deviation of area of the triangles	This is the standard deviation of the area of the triangles formed by using Delaunay triangulation on the cells detected.
F124	Average edge length	This is the mean of the length of the edges of the triangles formed.
F125	Std deviation of edge length	Standard deviation of the length of the edges of the triangles formed.

Table 7.1. Feature table,(a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (k) Basal (cont.)

(k) BASAL FEATURES

F126	number of nuclei in the basal part/number of nuclei in the non-basal part of the epithelium	The higher the ratio it signifies advanced stages of CIN since the basal membrane in that case will have higher number of nuclei compared to the non-basal part.
F127	no. of nuclei in the basal part/area of the entire epithelium	This is the feature which helps us to estimate the nuclei concentration with respective to the image area. The higher the ratio more advanced the stage of CIN is for a given image.
F128	darkness of the basal part (color ratio precisely add up the gray level histogram value of the basal membrane)/ luminosity of the rest of the non-basal epithelium (color contrast)	This is a color ratio and measures how dark the basal membrane is with respect to the rest of the image. More the number of nuclei in the basal membrane more darker it will be and higher stages of CIN like CIN2 or CIN 3 can be expected
F129	eccentricity measure of the nuclei in the basal part/ sq. Root of the basal epithelium	Not a very pronounced feature but as CIN stages advances the nuclei tends to become more elliptical rather than circular.
F130	area(area of the epithelium populated by nuclei ONLY)/ sq. Root of the area of the entire epithelium	This ratio measures the nuclei area as compared to the sq. root of the area of the entire epithelium. An again higher ratio shows advanced stages of CIN.
F131	Percentage of the area of the basal epithelium to the entire epithelium. (Distance measure)	This precisely measures the width of the basal membrane. With advanced stages the basal membrane becomes wider as compared to early stages of the CIN.
F132	Average luminance basal /entire epithelium.	This average luminance of a CIN3 would be lower as compared to a CIN1 or normal image. So in this case the lower the ratio the higher the stage of the CIN is.

Table 7.1. Feature table,(a) texture, (b) color, (c) triangle, (d) WDD, (e) nuclei, (f) light area, (g) combined, (h) layer-by-layer triangle (m) Basal (cont.)

F133-F137	The difference in nuclei concentration measure as seen in strip1-strip10, strip2-strip9, strip3-strip8, strip4-strip7, strip5-strip6.	For advanced stages of CIN this ratio would vary very less as opposed to lower stages of CIN. This is because with a CIN3 the basal membrane is spread for almost 60-90% of the image as opposed to a 10 or 20 % normal or CIN1 image.

7.1 NUCLEI FEATURE DEVELOPMENT

Nuclei feature development detects the nuclei both location wise and count wise on the epithelium region that estimates the classification of CIN stage. The algorithm was developed by Lu [21] and Guo [23], including two features calculated within vertical segments as shown below:

1. Number of nuclei

2. Nuclei area over background area

The steps for computing the nuclei-based features are as follows. First, with the nuclei detected, in an image called nuclei mask, the number of nuclei can be counted, also the nuclei area can be found.

Secondly, the epithelium background without the nuclei are computed with the help of nuclei mask like entire epithelium – nuclei mask, and the area of background is calculated.

7.2 LIGHT AREA FEATURE DEVELOPMENT

For each vertical segment within the epithelium, to define classification by light areas, some relative feature data were extracted. The following Figure 7.1 shows some light areas.

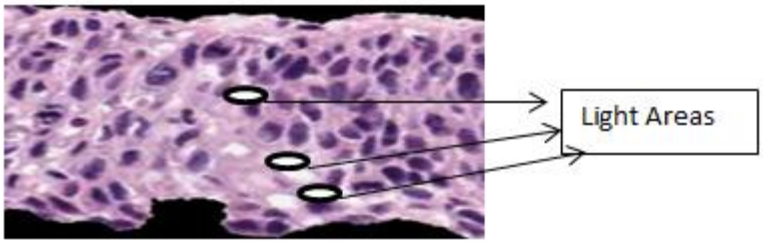


Figure 7. 1. light areas.

Upon thorough investigation it was observed that individual analysis of the R,G and B plane of an RGB image with respect to color concentration analysis and the L plane obtained after converting the color model to La*b* space was most beneficial. Since the light areas are not “white” strictly speaking at times they tend to merge with the epithelium. Individual analysis of the above mentioned color planes have shown to yield better results than using any other color space. The computed ratios are namely ratio RGB, ratio R, ratio G, ratio B, and ratio of luminance image in L plane, number of light area in per area and the ratio of light area to background area separately. The mask creation follows from section 5.2. The exact procedures are as follow:

1. Determine the mask of light area, RGB segmented image using the light-area mask and luminance image. The detailed process of the light area mask determination is explained in section 5.2 LIGHT AREA DETERMINATION. The mask

obtained by this process is binary but was converted to several other data types for respectively applying it to individual R, G, B and L plane.

2. Calculate the average intensity of the segmented RGB image and the associated average background intensity. Average intensity is basically calculated with the help of histogram analysis and finding out the mean ((R+G+B)/3).

3. Apply the same algorithm of histogram and mean to get the average intensity of R-plane, G-plane, B-plane, L-plane in luminance image and the associated average background intensity.

4. Obtain the total number and final areas of the light parts, and calculate the total area of the whole image.

5. Divide the average intensity of RGB image by its background area which is the area except white area to obtain the ratio of RGB.

6. The same calculation as mentioned in step 5 was used to get ratio of R, ratio of G, ratio of B, ratio of luminance.

7. Divide the total number of light parts by total area of the whole epithelium image to get the number of white area in per unit area.

8. Compute the ratio of light areas to background areas. One comparison is light-area by non-light areas and the other one is light-area by entire image.

7.3 LAYER-BY-LAYER TRIANGLE FEATURES

The layer-by-layer triangle algorithm and the features generated are presented here are developed by Guo [23], taking advantages the concept of “Delaunay Triangle” and is an extension of previously derived Triangular features by De [8]. In this algorithm,

the nuclei detection result is imported to locate the position of the nuclei which are used as vertices forming the Delaunay triangles. In previous research [8], triangle features were developed and applied over the whole epithelium region. Nuclei were determined using the Hough Transform, and the centroids of the nuclei were used as triangle vertices. The features that are obtained from the triangles include: average area of the triangles, standard deviation of the area of the triangles, average distance between the vertices of the triangles found and standard deviation of the distance between the vertices of the triangles.

For the layer-by-layer features, the whole epithelium is divided into 3 equal size horizontal segments. The distribution of nuclei are quantized in each of the three layers to determine the extent that the nuclei have spread across the epithelium, which is an important attribute for characterizing the different CIN grades. Processing the nuclei detection result is shown in Figure 7.2, where the circles in three colors mark the centroids of nuclei in three different layers. The middle figure shows the nuclei mask for the given epithelium and the rightmost figure shows the nuclei centers marked in color belonging to any of the 3 horizontal layers.

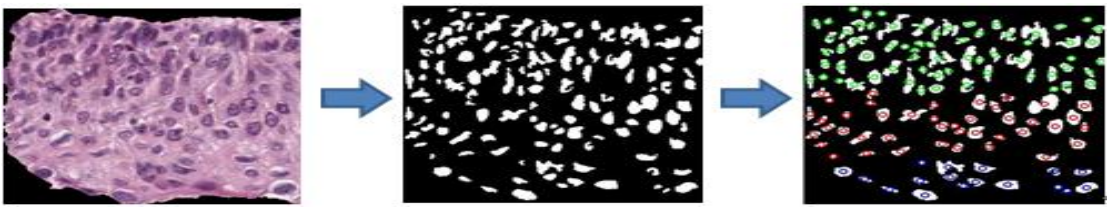


Figure 7. 2. Progress of locating nuclei (vertex) in three different layers.

Then, Delaunay triangles are determined using the centroids of the nuclei. The Delaunay triangles for the three different layers are shown in an example in Figure 7.3.

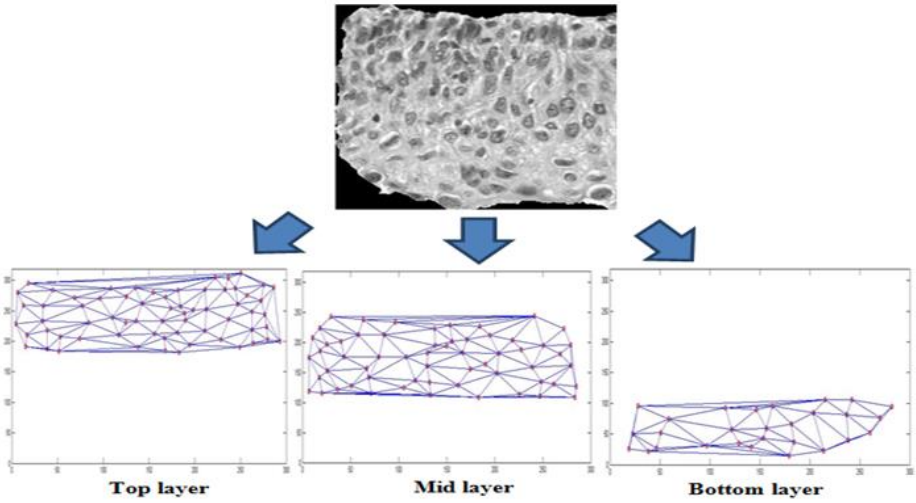


Figure 7. 3. Distribution of triangles for the entire image.

The top/bottom orientation of the epithelium is determined so that the higher density level of nuclei is top and the lower density is bottom, so that all the features obtained from calculation can be in a same order, which makes them useful and comparable.

Then, different features are calculated based on three different layers, including: the number of triangles in each layer, average area of the triangles in each layer, average edge distance of triangles in each layer, and the standard deviation of the two former features. For these triangle features mentioned, which describe different aspects of CIN stages, are all generalized by dividing the square root of epithelium area in the single vertical segment. The number of triangles can be another way of showing the number of nuclei existing per unit area, and it could be evidence of current CIN stage since the higher the CIN stage goes, relatively more nuclei would be there in per unit area of epithelium region. And the same thing happens on the average area of the triangles which

shows the number of triangles existing per unit area of epithelium region. Also the information of nuclei density is reflected on the feature value of average edge distance and standard deviation of triangles, as the higherthe average edge distance be, the lower the density would be. Figure 7.4 presents an example image where the lines represent the edges of the triangles and the vertices represent the positions of nuclei found and the different colors show the different layers.

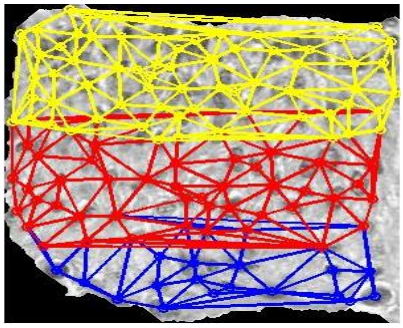


Figure 7. 4. Distribution of triangles in a single image.

7.4 BASAL MEMBRANE FEATURES

Once the basal membrane has been detected there are a number of ratios that has been extracted/calculated from it. The following describes them along with the reason/ justification in considering the mentioned ratios. Every ratio has verbal definition along with an equation defined from the feature table in Table 7.1.

$F_{126=}\ N_b/N_{nb}$ Where,

$N_b = Number\ of\ nuclei\ in\ the\ basal\ part$ And

$N_{nb} =\ number\ of\ nuclei\ in\ the\ non-basal\ part\ of\ the\ epithelium.$

The rationale for this feature is that higher ratios are intended to be associated with higher CIN grades since the basal membrane in that case will have higher number of nuclei compared to the non-basal part.

$F_{127=}\ N_b/A_E$, where

$N_b = Number.\ of\ nuclei\ in\ the\ basal\ part\ and,$

$A_E = Area\ of\ the\ entire\ epithelium.$

The rationale for this feature is that the feature provides an estimate of the nuclei concentration with respective to the image area. Higher ratios tend to be associated with higher CIN grades.

$F_{128=}\ D_b/L_{nb}\ Where,$

$D_b = Darkness\ of\ the\ basal\ part\ or\ luminesce\ values\ (from\ the\ gray\ level\ histogram\ of\ the\ basal\ membrane,\ adding\ gray\ level\ times\ its\ frequency)$

$L_{nb} = Luminosity\ of\ the\ rest\ of\ the\ non\text{-}basal\ epithelium\ (color\ contrast).$

Rationale for this feature is that this color ratio measures how dark the basal membrane is with respect to the rest of the image. More the number of nuclei in the basal membrane more dark it will be and higher stages of CIN like CIN2 or CIN 3 can be expected

$F_{129=}\ E_{nb}/As_b\ Where,$

$E_{nb} = Eccentricity\ measure\ of\ the\ nuclei\ in\ the\ basal\ part$

$As_b = \ Square\ root\ of\ the\ basal\ epithelium\ area$

Although this is not a very pronounced feature but as CIN stages advances the nuclei tends to become more elliptical rather than circular. The Eccentricity if defined

from the Basal mask. The idea to normalize the epithelium features with respect to square root was influenced by less computational burden and since the square root of a number is in between 0 and half of the number, at the worst case it's to compare numbers within the range of 1-9 as observed by trial and analysis which is acceptable in a neural network and is way faster.

$F_{130=}\ A_{N_b+N_{nb}}/As_{b+nb}$ Where,

$A_{N_b+N_{nb}}$ = *Average size of individual nuclei in the basal + non-basal part (area of the epithelium populated by nuclei ONLY)*

As_{b+nb} = *Square. Root of the area of the entire epithelium.*

Rationale for this feature ratio is that it measures the nuclei area as compared to the square root of the area of the entire epithelium. Again higher ratio shows advanced stages of CIN.

$F_{131=}\ A_{nb}/A_{b+nb}$ Where,

A_{nb} = *The area of the basal epithelium*

$A_{b+nb} = area\ of\ the\ entire\ epithilium$

Rationale for this feature mostly is that with advanced stages the basal membrane becomes wider as compared to early stages of the CIN.

$F_{132=}\ D_b/L_{n+nb}$ Where,

D_b = *Average luminance of the basal membrane (darkness measure)*

L_{n+nb} = *Average luminance of the entire epithelium.*

Rationale for this feature is for advanced stages of cancer this ratio would differ very less as opposed to lower stages of CIN. This is because with a CIN3 the basal membrane is spread for almost 60-90% of the image as opposed to a 10 or 20 % normal or CIN1 image.

$F_{133\ldots}F_{137=}\ segment_1\_segment_{10}, segment_2\_segment_9 \ldots segment_5\_segment_6$

Where, $segment_{i=}\ N_{ni}/A_i$ *, where, i=1,2,3…10*

$N_{ni} = The\ total\ number\ of\ nuclei\ in\ each\ of\ the\ strips, where\ we\ divide\ the\ entire\ epithilium\ into\ 10\ horizontal\ strips\ with\ strip1\ being\ the\ bottommost\ and\ Strip\ 10\ being\ the\ topmost.$

$A_i = area\ of\ the\ entire\ epithilium$

The rationale for this feature is to quantize the apparent relationship that images with higher CIN grades will have a lower ratio than for images with CIN grades. This is because with a CIN3 the basal membrane is spread for almost 60%-90% of the image as opposed to a 10 or 20 % normal or CIN1 image.

7.5 COMBINED FEATURE DEVELOPMENT

Combined features were developed to indicate the condition of CIN stage with respect to both nuclei and light area features. The features under investigation here are namely 'total number of Light areas/ total no of nuclei' and 'total area of the epithelium covered by light area/ total epithelium area covered by nuclei'. These combined features together are intended to provide discrimination among different CIN grades. The following equations help relate the ratios to Table 7.1.

$$F_{76} = \frac{F_{74}}{F_{67}}$$

Where, F76 represents $total\ number\ of\ Light\ areas\ /\ total\ no\ of\ nuclei$. The feature F74 is given as total number of light areas obtained over the epithilium mass and F67 is the total number of nuclei found over the same epithelium mass.

$$F_{77} = \frac{F_{75}}{F_{68}}$$

Where F77 denotes $total\ area\ of\ the\ epithelium\ covered\ by\ light\ area/\ total\ epithelium\ area\ covered\ by\ nuclei$. F75 yeilds the sum total area covered by all the light area features found in F74. F68, similarly is the sum total of the nuclei area from all the nuclei found in F67.

8. CLASSIFICATION

8.1 INDIVIDUAL VERTICAL SEGMENT CLASSIFICATION

As previously stated sixty-two cervix histology images with expert pathologist manual segmentations of the squamous epithelium and labeled CIN grades (16 Normal, 13 CIN1, 14 CIN2, and 18 CIN3) were obtained from NLM and examined in this study. For classifier training, the individual vertical segments for each image were assigned the CIN grades that the expert pathologist labeled the entire image. For example, if the expert pathologist labeled an image as CIN 1, all of the vertical segments for that image were designated as CIN 1. Experiments were performed partitioning the image into five and ten vertical segments. All experiments score CIN classification results based on the expert pathologist CIN grading of the images. For each vertical segment, 137 were extracted, as discussed in section 7. These features were used as inputs to a Particle Swarm Optimization neural network classifier to generate confidence values for the individual vertical segments, which are combined for image classification.

Evolutionary algorithms (EAs) using particle swarm optimization (PSO) for artificial neural network (ANN) training were investigated for individual vertical segment discrimination to facilitate image-based classification. PSO is the study of swarms of social organisms such as a flock of birds, in which each particle in the swarm moves toward its previous best location (Pbest) and global best location (Gbest) at each time step [30]. The PSO algorithm utilized in this study is presented in detail in [30] and is overviewed as follows.

The neural networks were trained using a leave-one-image out approach. For 10 vertical segments per image for 61 images gives 610 training feature vectors with

individual vertical segment label assignments) and 10 vertical segments for the left out image. In order to generate a continuum of values to represent the classes normal, CIN1, CIN2, CIN3, the target outputs for each vertical segment used in neural network training were assigned as 0, 0.33, 0.66, 1, respectively.

The PSO algorithm utilized in this study is presented in detail in [30] and is overviewed as follows. M particles (ANNs) are initialized with an architecture of D x R x 1, where D is the input feature, R is the number of hidden nodes, and one output. The connection weights in each ANN are updated when the elements in each particle are trained as follows. The initial value for each element of the vector is randomly set at a value from –0.1 to 0.1. In each training time step, the element's value of each particle is updated toward Pbest and Gbest. Pbest is a particle of the M particles that gives the least root mean square error (RMSE) between the current training epoch and the previous training epoch. Gbest is the particle among the M particles which generates the minimum RMSE, where the RMSE is calculated based on the difference between the ground truth and the actual ANN's output. The details for the updating process are given in [30]. The same process is repeated for N epochs. The final Gbest particle is selected for the final ANN weights for the test vector. The following Figure 8.1 shows the basic concept of PSO followed by the base equations.

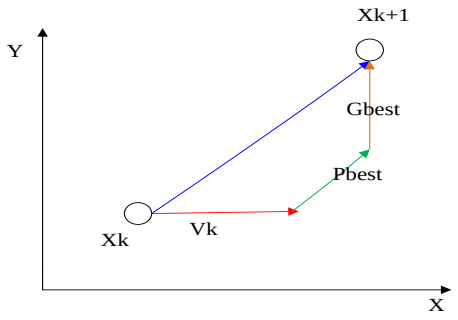


Figure 8. 1. Basic concept of PSO.

The basic PSO algorithm uses a real-valued multidimensional space as belief space, and evolves the position of each particle in that space using the following equations [30]:

$$v_{id}^{t+1} = w \cdot v_{id}^{t} + c_1 \cdot \psi_1 \cdot (p_{id}^{t} - x_{id}^{t}) + c_2 \cdot \psi_2 \cdot (p_{gd}^{t} - x_{id}^{t}) \quad \text{.....(1)}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad \text{..... (2)}$$

v_{id}^{t} : Component in dimension d of the i^{th} particle velocity in iteration t.

x_{id}^{t} : Component in dimension d of the i^{th} particle position in iteration t.

c_1 , c_2 : Constant weight factors.

p_i : Best position achieved so long by particle i.

p_g : Best position found by the neighbors of particle i.

ψ_1 , ψ_2 : Random factors in the [0, 1] interval.

w : Inertia weight.

The particle used to calculate p_g depends on the type of neighborhood selected. In the basic algorithm either a global (*gbest*) or local (*lbest*) neighborhood is used. In the global neighborhood, all the particles are considered when calculating p_g. In the case of the local neighborhood, neighborhood is only composed by a certain number of particles among the whole population. The local neighborhood of a given particle does not change during the iteration of the algorithm.

A constraint (v_{max}) is imposed on v_{id}^{t} to ensure convergence. Its value is usually kept within the interval [$-x_{id}^{max}$,x_{id}^{max}], being x_{id}^{max} the maximum value for the particle position. A large inertia weight (w) favors global search, while a small inertia weight favors local search. If inertia is used, it is sometimes decreased linearly during the iteration of the algorithm, starting at an initial value close to 1. An alternative formulation of Eq. 1 adds a constriction coefficient that replaces the velocity constraint (v_{max}). The PSO algorithm requires tuning of some parameters: the individual and sociality weights (c_1,c_2), and the inertia factor (w).

For the case of ten vertical segments determined from each image for the training set of images and the left out image, neural network outputs are determined. A weighted sum approach for combining neural network outputs was examined for combining the neural network outputs of the vertical segments for each image. Let Seg_1, Seg_2…, Seg_{10} denote the vertical segments for the 10 vertical segment decomposition of each image,. Let N_1, N_2,…, N_{10} denote the neural network outputs for each of the vertical segments. Let W_1, W_2, W_3, W_4, and W_5 denote the weights applied to the different vertical segments. Note that the weights are specified such that the segments at corresponding positions along the medial axis are given equal weights in order to accommodate for rotational variations (flipped or not flipped) in the way that the epithelium region is processed. The final output used for each image for the 10 vertical segment case for image-based classification is given as the equation and Table 8.1 below:

$$O_{img}^{10} = \sum_{i=1}^{5} (W_i N_i + W_i N_{11-i})$$

Table 8. 1. Input variables for each single segment among 10 vertical segments.

Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6	Seg 7	Seg 8	Seg 9	Seg 10
N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9	N_{10}
W_1	W_2	W_3	W_4	W_5	W_5	W_4	W_3	W_2	W_1

The values of O_{img}^{10} for image-based CIN grade classification, which will be presented in more detail in section 9.

A similar process was explored for five vertical segments computed from each image for the training set of images and for the left out test image. Let Seg_1, Seg_2, Seg_3, Seg_4, and Seg_5 denote the five vertical segments determined from an image. Let N_1, N_2, N_3, N_4, and N_5 denote the neural network outputs for the five vertical segments, respectively, which are shown in Table 8.2. In comparison to 10 segments the 3 segments are treat as if Seg 1 and Seg 2 of the 10 Segments fuse together to give Seg 1 of the 5 segments and so on. Let W_1, W_2, and W_3 denote the weights applied to the different vertical segments, with the weights specified so that segments at corresponding positions along the medial axis are given equal weights in order to accommodate for rotational variations (flipped or not flipped) in the way that the epithelium region is processed.

Table 8. 2. Input variables for each single segment among 5 vertical segments.

Seg 1	Seg 2	Seg 3	Seg 4	Seg 5
N_1	N_2	N_3	N_4	N_5
W_1	W_2	W_3	W_4	W_5

The final output used for each image for the 5 vertical segment case for image-based classification is given as:

$$O_{img}^{5} = \sum_{i=1}^{5} W_i N_i$$

The values of O_{img}^{5} for image-based CIN grade classification, which will be presented in more detail in section 9.1.

9. EXPERIMENTS PERFORMED

We performed four sets of experiments, which are presented in this section.

9.1 IMAGE-BASED WEIGHTED SUM CONFIDENCE VALUE DETERMINATION

Features (see section 7) were extracted from each vertical segment for the 62 image data set. Using the leave-one-image out training and testing approach, the features extracted from the vertical segment images were used as inputs to train the PSO neural network classifier. The PSO neural network outputs were for each vertical segment from each training image and from the left out test image. Then, the weighted sum confidence values were computed from the PSO neural network outputs, O_{img}^{10} for ten vertical segments and O_{img}^{5} for five vertical segments, for each training and test image for image-based CIN grade classification.

For CIN image-based classification, different feature combinations for PSO neural network training and testing were investigated (see Table 7.1), including: texture features, color features, triangle features, basis function correlation features (WDD features), nuclei features, light area features, combined features, layer-by-layer triangle features, and Basal membrane features. Combinations of these feature groups were also examined for CIN grade image-based classification.

CIN grade assignment to each test image was performed as follows for the leave-one-out image approach. The weighted sum confidence values from the training set of images for each leave-one-out image case, the set of O_{img}^{10} from the training set of images for ten vertical segments and the set of O_{img}^{5}, from the training set of images for five vertical segments, were sorted to generate a classification continuum. From the training

set of images for PSO neural network training, expert truthed normal, CIN1, CIN2, and CIN3 images were designated as 0, 0.33, 0.66, and 1 value, which were assigned to all of the vertical segments for those images. Accordingly, performing a weighted sum of the PSO neural network outputs generated a sum, which was expected as low for normal images and increasing with severity of the CIN grade. Based on this premise, the weighted sum confidence values for the test set of images were sorted, and thresholds of confidence values were automatically determined which maximized the exact CIN grade classification rate (see Approach 1 below) for the test set of images. Exhaustive search of the different weight combinations is performed for determining the highest classification rate. Note that different weight combinations lead to the same classification rate (linear scaling). The resulting thresholds are applied to CIN grade classify the left out image. The process is repeated with all images left out once, and the individual image CIN grade classifications are compared to the expert CIN grade labels for scoring.

Four scoring approaches for evaluation of the epithelium image classifications were explored:

Approach 1 (Exact Class Label): The first approach is exact classification, meaning that if the class label automatically assigned to the test image is the same as the expert class label, then the image is considered to be correctly labeled. Otherwise, the image is considered to be incorrectly labeled.

Approach 2 (Windowed Class Label): The second scoring approach is a windowed classification scheme. Using this approach, if the predicted CIN grade level for the epithelium image is only one grade off as compared to the actual CIN grade, we

considered it as correct classification. For example, if CIN1 was classified as Normal or CIN 2, the result would be considered correct. If CIN1 was classified as CIN3, the result would be considered incorrect.

Approach 3 (Normal versus CIN): For the third approach, we considered the classification incorrect when a Normal stage was classified as any CIN stage and vice-versa.

Approach 4 (Normal and CIN1 versus CIN2 and CIN3): For the fourth approach, we considered the classification incorrect when a CIN2 or CIN3 grade was called a Normal or CIN1 grade and vice-versa. Figure 9.1below shows the classification scoring schemes for approaches 1-3.

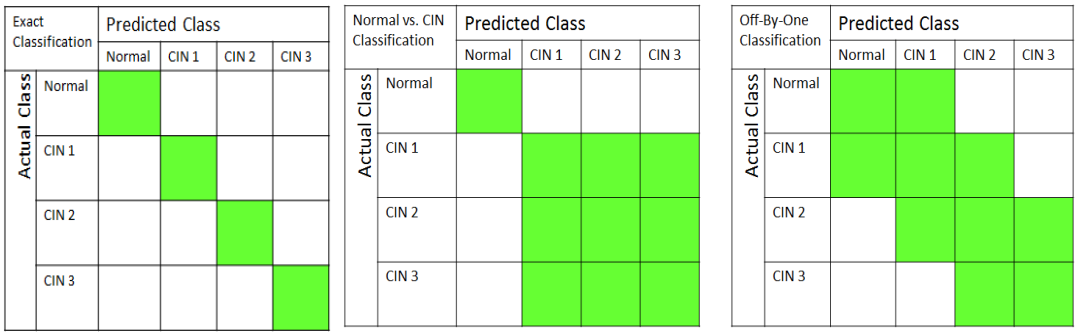


Figure 9. 1. Different scoring schemes for 3 kinds of classifications (a) Exact Classification (b) Normal Vs CIN classification (c) Off-By-One Classification.

The following figure presents a flowchart of the entire image-based CIN grade classification process below. Details about the Hybrid Neural network classifier and the weighted sum decision algorithm is presented in the following sections.

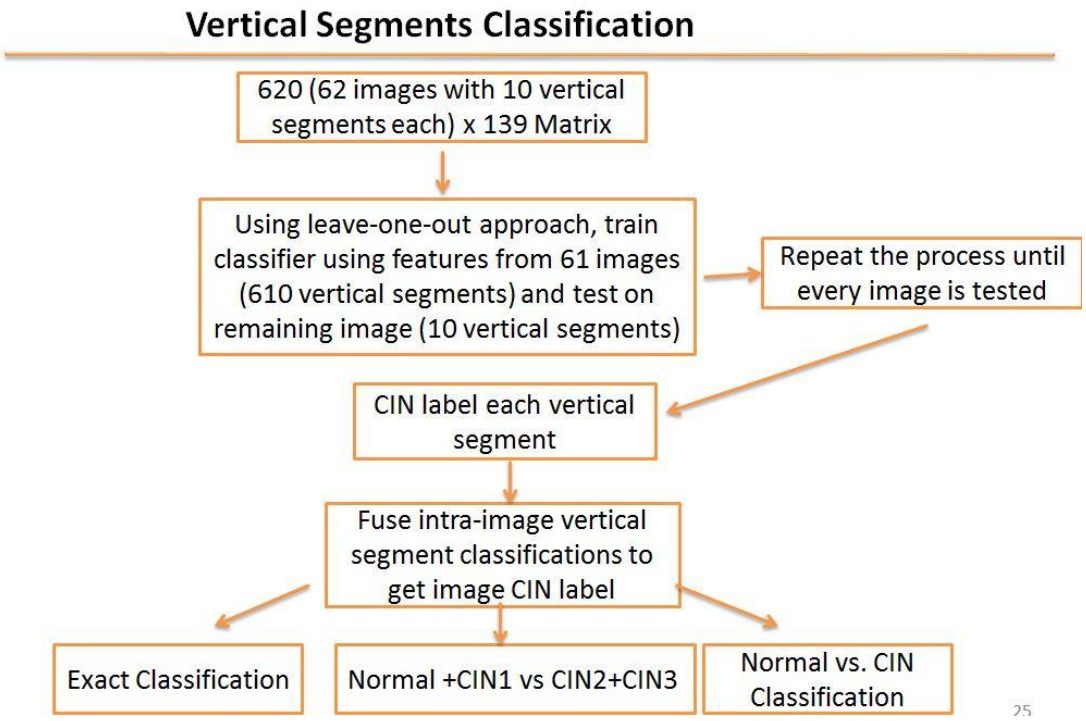


Figure 9. 2. Block diagram of vertical segments classification.

9.2 EXPERIMENTS PERFORMED AND EXPERIMENTAL RESULTS

The PSO neural network architecture with the weighted sum neural network output combination approach described above for the 10 and 5 vertical segment cases was examined for different feature combinations. Table 9.1 presents the exact class labels, the features used from Table 7.1 are listed in column 1 and 4 of Table 9.1 which give the different approaches for combining the neural network outputs for each of the vertical segments (10 vertical segments examined here). The analysis here is done with respect to exact class label and Normal vs. CIN scoring schemes. Later, all four scoring scheme results are reported for the few best combination groups.

Table 9. 1. Image-based classification results using PSO neural network approach for continuous classification scale. (Note that 10 vertical segments are used for feature analysis for each image).

Features Used	Exact Class Label	Normal vs. CIN	Features Used	Exact Class Label	Normal vs. CIN
Texture	69.35	83.87	Color, Nuclei, Light Area, Combined	72.58	93.55
Correlation WDD	66.13	93.55	Color, Layer-by-Layer Triangle	53.23	77.42
Nuclei, Light Area, Combined	74.19	93.55	Color, New Triangle 2, New Triangle 3	61.29	83.87
Layer-by-Layer Triangle	48.39	74.19	Color, Basal Membrane	61.29	83.87
New Triangle2	46.77	70.97	Nuclei, Light Area, Combined, Layer-by-Layer Triangle	69.35	90.32
New Triangle3	59.68	83.87	Texture, Color, Nuclei, Light Area, Combined	77.42	93.55
Basal Membrane	53.22	70.97	Texture, Color, Layer-by-Layer Triangle	77.42	96.77
Texture, Color	80.65	93.55	Texture, Color, Layer-by-Layer Triangle, Basal Membrane	74.19	93.55
Texture, Triangle	67.74	83.87	Color, Nuclei, Light Area, Combined	75.81	93.55
Texture, Correlation WDD	80.65	93.55	Correlation WDD, Nuclei, Light Area, Combined	67.74	90.32
Texture, Nuclei, Light Area, Combined	74.19	87.1	Texture, Color, Nuclei, Light Area, Combined	80.64	93.55
Texture, Layer-by-Layer Triangle	66.13	87.1	Texture, Color, Nuclei, Light Area, Combined, Combined WDD	79.03	93.55
Texture, New Triangle 2	64.52	83.87	Nuclei, Light Area, Combined, Layer-by-Layer Triangle, New Triangle 2, New Triangle 3, Basal Membrane	79.03	93.55
Texture, New Triangle 3	66.13	87.1	Texture, Color, Triangle, Nuclei, Light Area, Comb Layer-Triangle, Basal	83.87	96.77

Table 9.1. Image-based classification results using PSO neural network approach for continuous classification scale. (Note that 10 vertical segments are used for feature analysis for each image). (Cont.)

Texture, Basal Membrane	66.13	87.1	Texture, Color, Triangle, Nuclei, Light Area, Combined, Combined Layer-by-Layer, New Triangle 2, Basal Membrane	88.71	96.77
Color, Correlation WDD	59.68	90.32	Texture, Color, Nuclei, Light Area, Combined, Layer-by Layer Triangle, New Triangle 3, Basal Membrane	82.26	96.77

Finally, feature analysis based on Fisher's scoring optimization technique for stepwise variable selection in SAS was performed to determine statistically significant features. Probability values (Pr) is used with the Chi-Square scores for variable selection. Note that a binary model was examined here (Normal vs. CIN) for feature selection and statistical classification of the individual vertical segments. This particular feature group was chosen since it can tolerate maximum number of faults. Although not very accurate this was just an analysis.

\

Table 9. 2. Probability values (Pr) used for different features.

Feature	Pr value	Feature	Pr value	Feature	Pr value	Feature	Pr value	Feature	Pr value
F1	<0.0001	F13	<0.0001	F15	0.0348	F19	<0.0001	F67	0.0329
F2	0.0001	F14	<0.0001	F16	<0.0001	F21	0.577	F68	<0.0001
F4	0.0611			F17	0.0031	F23	<0.0001		
F8	<0.0001					F30	<0.0001		
F9	<0.0001					F42	0.0101		
F10	<0.0001					F44	0.0041		
						F56	0.0882		
						F58	0.0003		
						F61	<0.0001		
Individual Segment Classification on Normal Vs CIN	90.80%		87.90%		77.30%		86.30 %		91.80%

Feature	Pr value	Feature	Pr value	Feature	Pr value	Feature	Pr value
F70	<0.0001	F76	0.0006	F78	0.0005	F126	<0.0001
F71	<0.0001	F77	<0.0001	F79	0.0015	F127	0.0277
F72	<0.0001			F80	0.0734	F128	<0.0001
F74	0.0094			F83	<0.0001	F129	0.0002
F75	<0.0001			F89	0.0007	F130	0.0032
						F131	<0.0001
						F134	0.04

Table 9.2. Probability values (Pr) used for different features (cont.)

Individual Segment Classification (Normal vs. CIN)	92.30%		85.00%		80.50%		85.60%

Based on feature reduction and examining different feature groups as shown in Table 9.2 using 5 vertical segments for each epithelium region, the following feature combinations with classification results are presented in Table 9.3. Note that the experimental results are presented for Exact Class Label, Off-by-One Window, Normal versus CIN, and Normal and CIN1 versus CIN2 and CIN3 (3 total classes).

Table 9. 3. Classification results for 5segments with different feature combinations using Exact Class Label, Off-by-One Window, Normal vs. CIN, and Normal vs. CIN1 vs. CIN2 or CIN3.

Feature Combination	Exact Class Label	Off By One	Normal vs. CIN	Normal/ CIN1 vs CIN2/CIN 3	Segment Weights for Weighted Classifier Sum (W1,…,W5)
F1-F18, F67-F77, F124	90.32%	98.39%	96.77%	96.77%	1,0,0,0.75,0
F1-F18,F29,F30, F67-F77,F124, F126,F136	90.32%	98.39%	96.77%	96.77%	0.25,0.75,0.25,0.5,0.25

Extending the feature reduction and analysis from the 5 vertical segments, classification results for different feature groups based on 10 vertical segments are presented in Table 9.4. Again, note that the experimental results are presented for Exact

Class Label, Off-by-One Window, Normal versus CIN, and Normal versus CIN1 versus CIN2/CIN3 (3 total classes). Many different feature combinations were examined. The feature combinations presented in Table 9.4 provided the highest classification results.

Table 9. 4. Classification results for different feature combinations based on 10 vertical segments using Exact Class Label, Off-by-One Window, Normal vs. CIN, and Normal vs. CIN1 vs. CIN2/CIN3.

Feature Combination	Exact Class Label	Off By One	Normal vs. CIN	Normal vs CIN1 vs CIN2/CIN3	Segment Weights for Weighted Classifier Sum (W1,…,W5)
F1-F18, F67-F77	91.94%	100.00%	96.77%	96.77%	1,0,0,0.75,0
F1-F18, F67-F77,F124,F126,F136	88.71%	100.00%	96.77%	93.55%	1,0.5,0,1,1
F1-F18,F29,F30,F67-F77,F124,F126,F136	85.48%	100.00%	96.77%	90.32%	0.25,0,0.75, 0,0.25

Table 9.3 and Table 9.4 show actual image classification rates of 90.32% and 91.94% for the 5 and 10 vertical segment cases, respectively. The feature combinations that yielded the highest classification results for the 10 and 5 vertical segment cases include the texture features (F1-F11), color features (F12-F14), triangle features (F15-F18), nuclei features (F67,F68), light area features (F69-F75), and combined features (F76,F77).

9.3 INTER-PATHOLOGIST IMAGE-BASED CLASSIFICATION OF DIGITIZED CERVICAL IMAGE DATA SET

Dr. Rosemary Zuna, Professor of Pathology, at the University of Oklahoma Health Sciences Center provided the expert pathologist CIN grades for the 62 digitized histology images of the epithelium region presented in the previous sections of this thesis.

In this section and in following sections, a second expert pathologist was sought for guidance in establishing CIN truth classifications for sub-regions of the epithelium (individual vertical segments) and for the entire epithelium region. Dr. Shelly Frazier, Surgical Pathologist, at the University of Missouri was approached and agreed to provide these classifications for the 62 cervical images data set. Note that all vertical segment classifications used in this study were provided by Dr. Zuna and Dr. Frazier.

Overall, the CIN classification results from Table 9.5 before and Table 9.6 below are similar, with slightly higher classification results obtained based on the expert CIN labeling from Dr. Zuna. Using Dr. Frazier's CIN labeling of the 62 image data set with 10 and 5 vertical segments, several features set combinations were examined using different scoring schemes, including the Exact Class Label, Normal vs. CIN and Normal+CIN1 vs. CIN2+CIN3.

Table 9. 5. Classification results based on CIN truth grades from Dr. Frazier for different feature combinations based on 10 vertical segments using Exact Class Label, Off-by-One Window, Normal vs. CIN, and Normal vs. CIN1 vs. CIN2/CIN3.

Feature Combination	Actual Image	Off By One	Normal vs. CIN	Normal vs CIN1 vs CIN2/CIN3	Segment Weights for Weighted Classifier Sum (W1,…,W5)
F1-F18, F67-F77	88.71%	98.39%	96.77%	91.94%	1,0.25,1,0.25,0
F1-F18, F67-F77, F124,F126,F136	88.71%	98.39%	96.77%	93.55%	1,0.5,0,1,1

The individual 5 vertical segment CIN labels for classifier training were the image-based CIN labels provided by Dr. Frazier. Here the same PSO-based classifiers were used to train and test the images with leave one image out scheme. The following parameters were used in the PSO neural network algorithm: rg_vals = 0.001, rm_vals =

0.001, c2_vals = 1.5; c1_vals = 1, particles_number_vals = 30, w_vals = 0.6. Tables 8.8 and 8.9 present the classification results for the different features.

Table 9. 6. Classification results for different feature combinations with 10 vertical segments for Exact Class Label, Normal vs. CIN, and CIN1 vs. CIN2/CIN3 based on Dr. Frazier's CIN labeling.

Features Used	Exact Class Label (%)	Normal vs. CIN (%)	Normal + CIN1 vs. CIN2 + CIN 3 (%)	Features Used	Exact Class Label (%)	Normal vs. CIN (%)	Normal + CIN1 vs. CIN 2 + CIN3 (%)
Texture	69.35	83.87	83.87	Nuclei, Light Area, Combined, Layer-by-Layer Triangle	69.35	90.32	91.94
Texture, Color, Triangle, Nuclei, Light Area, Combined, Layer-by-Layer Triangle, Basal Membrane	88.71	96.77	93.55	Texture, Triangle	67.74	83.87	83.87
Texture, Color, Nuclei, Light , Combined, Layer-by Layer Triangle, Basal Membrane	82.26	96.77	93.55	Correlation WDD, Nuclei, Light Area, Combined	67.74	90.32	91.94
Texture, Color	80.65	93.55	91.94	Correlation WDD	66.13	93.55	93.55

Table 9.6. Classification results for different feature combinations with 10 vertical segments for Exact Class Label, Normal vs. CIN, and CIN1 vs. CIN2/CIN3 based on Dr. Frazier's CIN labeling (cont.)

Texture, Correlation WDD	80.65	93.55	93.55	Texture, New Triangle 3	66.13	87.1	83.87
Texture, Color, Nuclei, Light Area, Combined	80.64	93.55	95.16	Texture, Basal Membrane	66.13	87.1	82.26
Texture, Color, Nuclei, Light Area, Combined, Combined WDD	79.03	93.55	91.94	Texture, New Triangle 2	66.13	87.1	85.48
Nuclei, Light Area, Combined, Layer-by-Layer Triangle, New Triangle 2, New Triangle 3, Basal Membrane	79.03	93.55	91.94	Color, New Triangle 2, New Triangle 3	61.29	83.87	82.26
Color, Nuclei, Light Area, Combined	72.58	93.55	87.1	New Triangle2	46.77	70.97	72.58

Table 9. 7. Classification results for different feature combinations with 5 vertical segments for Exact Class Label, Normal vs. CIN, and CIN1 vs. CIN2/CIN3 based on Dr. Frazier's CIN labeling.

Features Used	Exact Class Label (%)	Normal vs. CIN (%)	Normal + CIN1 vs. CIN 2 + CIN 3 (%)	Features Used	Exact Class Label (%)	Normal vs. CIN (%)	Normal + CIN1 vs. CIN 2 + CIN3 (%)
Texture, Color, Nuclei, Light Area, Combined, Layer-by Layer Triangle, New Triangle 3, Basal Membrane	90.32	96.77	96.77	Texture, Color, Layer-by-Layer Triangle	83.87	93.55	91.94
Texture, Color, Nuclei, Light Area, Combined	88.71	96.77	96.77	Texture, Color, Nuclei, Light Area, Combined, Combined WDD	83.87	93.55	95.16
Texture, Color, Triangle, Nuclei, Light Area, Combined, Layer-by-Layer Triangle, Basal Membrane	88.71	96.77	96.77	Nuclei, Light Area, Combined, Layer-by-Layer Triangle, New Triangle 2, New Triangle 3, Basal Membrane	77.42	96.77	93.55
Texture, Color	87.1	93.55	96.77	Texture, Correlation WDD	74.19	90.32	87.1
Texture, Triangle	83.87	83.87	90.16				

From Table 9.6, the best 10 vertical segment classification results are 88.71% for the Exact Class Label, 95.16% for CIN vs. Normal and 93.55% for Normal+CIN 1 vs. CIN2 + CIN3. From Table 9.7, the best 5 vertical segment classification results are90.32% for the Exact Class Label, 96.77% for CIN vs. Normal and 96.77% for Normal + CIN1 vs. CIN2 + CIN3.

The combination of features which yielded the highest classification results contains texture, color, nuclei, light area, combined features, triangle, and only three sub-features in the group of basal membrane features. Texture and color feature provide the general structural analysis of the data. Nuclei, light area, and combined features, as well as triangle features give much information about the characters in different CIN stages, they are more dynamic and differ with the increasing of CIN stage with which a clue can be found in contributing to make a proper diagnosis.

The other group of features like Correlation WDD, Triangle features discussed in former paper, and most of the basal membrane do not yield as good a result as those mentioned above do. And, the basal membrane feature is a little different from other features but also follow the rules that features are operated in different layers, which finally give information about the origin of certain CIN stages. With some concern, these features fail to provide some “key information” which leads a significant difference in describing the current situation of the whole epithelium, or in other words, are not typical or universal for all the epithelium regions which are taken into the test. Some of the features are generated to be trials or control group, in order to discover more clues which can lead to a better result.

From the results presented in [8], feature combination of texture, original triangle and correlation WDD features yielded the exact label classification result of 70.5% and Normal vs. CIN result to be 90.2%. The results presented in this research had an exact label classification result of 90.32% and Normal versus CIN of 96.77%, yielding improvements of 28.11% and 7.28%, respectively.

10. CONCLUSION

In this study, an automated CIN grade classification of vertical segmented epithelium regions is developed. The method developed includes medial axis determination, bounding box determination and partitioning the whole epithelium region into several vertical segments with the respect of medial axis. Then as many as 137 features are generated and taken all through experiment procedures which include leave-one out, normal vs. CIN, Normal+CIN1 vs. CIN2+3, off by one, and yield the final results through data fusion. And the features generated consist of texture, color, triangle, nuclei, light area, combined features, layer-by-layer triangle features, and basal membrane features.

Experimental results from this study show higher CIN classification with 90.32% for exact label classification and 96.77% for normal vs. CIN classification compared to 70.5% and 90.2%, respectively, from previous research [8]. There has been a significant improvement over the classification results as compared to initially reported results in [8], which used an SVM-based classifier for image-based classification. The improvement in the current research is mainly contributed from the developed new features and the PSO neural network and weight sum approach for individual vertical segment conference value determination and fusion for image-based classification. Some of the features in this study such as nuclei, light area, and layer-by-layer triangle features outperform other features and contribute in improving the CIN classification, demonstrating the potential for vertical segmentation and the horizontal layer by layer analysis for enhancing CIN grading for the epithelium. Overall, most of the CIN grade assessments for the epithelium histology images from the two pathologists agree

contribute to the similar CIN classification results reported in this thesis. The epithelium images which the pathologists disagree provide the basis for discovering different ways to fuse classification data for each single segment or a different method to update the weight in the neural network used in classification.

BIBLIOGRAPHY

[1] Parkin DM, Bray FI, Devesa SS. Cancer burden in the year 2000: the global picture. Eur J Cancer 2001; 37 (Suppl. 8): pp 54-66.

[2] Jeronimo J, Schiffman M. A tool for collection of region based data from uterine cervix images for correlation of visual and clinical variables related to cervical neoplasia. In: Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems. 2004. pp. 558-62

[3] Kumar V, Abbas A, Fausto N, Cotran R. pathologic basis of disease. 8th ed. Philadelphia (PA): Saunders Elsevier; 2009.

[4] He L, Long LR, Antani S, Thoma GR. Computer assisted diagnosis in histo-pathology. In: editor. Sequence and genome analysis: methods and applications. iConcept Press; 2001. pp. 271-87.

[5] Wang Y, Crookes D, Eldin OS, Wang S, Hamilton P, Diamond J. Assisted diagnosis of cervial intraepithelial neoplasia (CIN). IEEE J Sel Topics Signal Process 2009; 3(1): pp. 112-21.

[6] McCluggage WG, et al. Inter- and intra-observer variation in the histopatho-logical reporting of cervical squamous intraepithelial lesions using a modified Bethesda grading system. Int J Obstet Gynecol 1998; 105(2): pp. 206-10.

[7] Ismail SM, Colclough AB, Dinnen JS, Eakins D, Evans DM, Gradwell E, O'Sullivan JP, Summerell JM, Newcombe R. Reporting cervial intra-epithelial neoplasia agreement. Histopathology 1990; 16(4): pp. 371-6.

[8] De S, Stanley RJ, Lu C, Long R, Antani S, Thoma G, Zuna R. A fusion-based approach for uterine cervical cancer histology image classification. Comput Med Imaging Graph 2013; 37: pp. 475-87.

[9] Molloy C, Dunton C, Edmonds P, Cunnane MF, Jenkins T. Evaluation of colposcopically directed cervical biopsies yielding a histologic diagnosis of CIN 1, 2. J Lower Genital Tract Dis 794 2002;6(2): pp. 80–3.

[10] Soenksen D. Digital pathology at the crossroads of major health care trends: corporate innovation as an engine for change. Arch Pathol Lab Med 2009;133: pp. 555–9.

[11] Keenan SJ, Diamond J, McCluggage WG, Bharucha H, Thompson D, Bartels PH, Hamilton PW. An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN). J Pathol 2000;192(3): pp. 351–62.

[12] Loménie N, Racoceanu D. Point set morphological filtering and semantic spatial configuration modeling: application to microscopic image and bio-structure analysis. Pattern Recogn 2012; 45(8):28 pp. 94–911.

[13] Guillaud M, Cox D, Malpica A, Staerkel G, Matisic J, Niekirk DV, Adler-Storthz K, Poulin N, Follen M, MacAulay C. Quantitative histopathological analysis of cervical intra-epithelial neoplasia sections: methodological issues. Cell Oncol 2004;26: pp. 31–43.

[14] Guillaud M, Adler-Storthz K, Malpica A, Staerkel G, Matisic J, Niekirkd DV, Cox D, Poulina N, Follen M, MacAulay C. Subvisual chromatin changes in cervical epithelium measured by texture image analysis and correlated with HPV. Gynecol Oncol 2005;99:S16–23.

[15] Miranda GHB, Soares EG, Barrera J, Felipe JC. Method to support diagnosis of cervical intraepithelial neoplasia (CIN) based on structural analysis of histological images. In: Proceedings of the 25th International Symposium on Computer-Based Medical Systems (CBMS). 2012. pp. 1–6.

[16] Price GJ, Mccluggage WG, Morrison ML, Mcclean G, Venkatraman L, Diamond J, Bharucha H, Montironi R, Bartels PH, Thompson D, Hamilton PW. Computerized diagnostic decision support system for the classification of preinvasive cervical squamous lesions. Hum Pathol 2003; 34(11): pp. 1193–203.

[17] Rahmadwati R, Naghdy G, Ros M, Todd C. Computer aided decision support system for cervical cancer classification. In: Proceedings of SPIE, Applications of Digital Image Processing XXXV, 8499. 2012. pp. 1–13.

[18] Wang Y, Chang SC, Wu LW, Tsai ST, Sun YN. A color-based approach for automated segmentation in tumor tissue classification. In: Proceedings of the Conference of IEEE Engineering in Medicine and Biology Society. 2007. pp. 6577–80.

[19] Wang Y, Turner R, Crookes D, Diamond J, Hamilton P. Investigation of methodologies for the segmentation of squamous epithelium from cervical histological virtual slides. In: Proceedings of the International Machine Vision and Image Processing Conference. 2007. pp. 83–90.

[20] Marel JVD, Quint WGV, Schiffman M, van-de-Sandt MM, Zuna RE, Terence-Dunn S, Smith K, Mathews CA, Gold MA, Walker J, Wentzensen N. Molecular mapping of high-grade cervical intraepithelial neoplasia shows etiological dominance of HPV16. Int J Cancer 2012;131:E pp. 946–53

[21] Lu C. Uterine cervical cancer histology image feature extraction and classification, M.S. Thesis, Missouri University of Science and Technology, 2013.

[22] De S, Data Fusion Techniques for Structural Health Monitoring and Signal Integrity Applications, Ph.D. Dissertation, Missouri University of Science and Technology, 2012.

[23] Guo P. Cervical cancer histology image feature extraction and classification, M.S. Thesis, Missouri University of Science and Technology, 2014.

[24] Maurer CR, Rensheng Q, Raghavan V. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. IEEE Trans Pattern Anal Mach Intell 2003;25(2): pp. 265–70.

[25] Soh LK, Tsatsoulis C. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Trans Geosci Remote Sens 1999;37(2): pp. 80–95.

[26] Stanley RJ, De S, Demner-Fushman D, Antani S, Thoma GR. An image feature-based approach to automatically find images for application to clinical decision support. Comput Med Imaging Graph 2011;35(5):3 pp. 65–72.

[27] Borovicka J. Circle detection using hough transforms. Course Project: COMS30121-image processing and computer vision. Technical report. University of Bristol; 2003. pp. 1–27.

[28] Rao CR, Toutenburg H, Fieger A, Heumann C, Nittner T, Scheid S. Linear models: least squares and alternatives. Berlin, Germany: Springer; 1999.

[29] Gonzalez R, Woods R. Digital image processing. 2nd ed. Englewood Cliffs (NJ): Prentice-Hall; 2002.

[30] Guidse G, Venayagamoorthy GK. Comparison of Particle swarm optimization and backpropagation as training algorithms for neural networks. Swarm Intelligence symposium 2003, pp. 110-117.

.

VITA

Koyel Banerjee was born in Durgapur in the State of West Bengal, India in 1989. She did her schooling at the Carmel Convent High School (1993-2005) and Hemsheela Model School (2005-2007) before going to West Bengal University of Technology for her Bachelor of Technology degree in Electronics and Communications Engineering from the Department of Electronics and Communications (2011). She stayed at Missouri University of Science and Technology, where she received her Master of Science degree in Computer Engineering from the Department of Electrical and Computer Engineering in December 2014.