CHAPTER 4

CLASSIFICATION OF GLANDULAR CELLS USING MULTIPLE COLOR SPACES

4.1 INTRODUCTION

In developed and developing countries, lung cancer is a major threat to mankind. One in four of all diagnosed involve lung cancer. The lung cancer remains the most common cancer-related cause of death both in developed and developing countries due to inhaling cancer-causing substances such as tobacco. Several methods such as MRI, chest- X rays, CT scan etc., are available for screening tests. For developing countries, the cost involved for early detection with the available method is not affordable. In this work, a novel low cost method to detect and classify lung glandular cells as Benign or Malignant (cancer cells) using pap stained sputum cytology images is given. The microscopic sputum images are preprocessed and analysis is restricted to cellular regions. For segmentation multiple color spaces and clustering algorithms such as K-means and Fuzzy C-means are used. Features are extracted and then Support vector machine are used for classification.

4.2 DATA SET

The sputum cytology images used in this work were provided by Pathology department, Regional Cancer Center Thiruvananthapuram, India. The nature of cells was confirmed by expert cytopathologist. The images were
captured by using a trinocular microscope fitted with a digital camera. The images were captured at 40X magnification level. The images are of 24 bit color depth with a spacial resolution of 3264x2448 pixels. The pap stained sputum samples are digitized using a digital microscope. The staining of the cellular materials varies widely in conventional pap staining. Normally the staining color varies from slide to slide and depends on various parameters like batch of stain prepared, age of cell, type of cell and also depends on the person preparing the slide. In this work, 1038 cell images are used which are cropped from the input sputum cytology images. A sample input cytology image is shown in the Figure 4.1.

![Figure 4.1 A sample input cytology image](image)

**4.3 SYSTEM ARCHITECTURE**

The system architecture for the proposed method to analyze the sputum cytology cells is shown in Figure 4.2. It includes preprocessing and segmentation of Pap stained sputum cytology images followed by feature extraction and classification of cell nuclei.
4.4 INPUT IMAGE

The sputum samples taken by cytotechnologists are preserved so that the cellularity of the nuclei is not lost. Then slides are prepared and stained using Pap stain. This is a differential stain in which the nuclear and cytoplasmic regions are stained in different colors. Nuclei region being more with dense materials and so will absorb more stain and appear much darker in color. The stained slide is placed under a trinocular light microscope and images are taken. The digital camera produces a high resolution image of 8 mega pixel (3264 X 2448) size. The slides are magnified at 40 X zoom level and images are taken. The images are stored on the workstation in 24 bit RGB color format. Images of both malignant and benign slides are taken. The classification of the slides as benign and malignant is done by experienced cytopathologists.
Step 2 : Preprocessing step: High resolution input image is downsized to 1/8th of the original size and use Blom’s method to find local minima. Use Threshold to locate the nuclear region.

Step 3 : Segmentation using clustering Techniques like K-Means clustering and Fuzzy C-Means clustering on Various Color spaces such as HSV, CIELAB, CIEXYy and CIELUV.

Step 4 : Features are extracted using catastrophe points

Step 5 : Classification Using SVM

Step 6 : Results

The steps for preprocessing as follows:

**4.5 PREPROCESSING**

The algorithm for preprocessing step is as follows:

Preprocessing : Algorithm

Step 1 : Input : Image of 8 Megapixel (3264 * 2448)

Step 2 : Resize the Image of Size 408 * 306

Step 3 : Find Local minima of Gaussian smoothed Image

Step 4 : Map co-ordinate of all minima to original Image

Step 5 : Crop a region of 128 * 128 size from the surroundings of all local Minima

Step 6 : Save the image if average intensity is less than 200 and Standard Deviation is greater than 30
The nuclei regions present in the image appears as dark blobs due to the absorption of stain in comparatively higher level than the cytoplasmic regions. By finding the local minima of the blobs the location of nuclei can be obtained. The high resolution input image is very difficult to process as such, so in this stage the image is down sized to 1/8\textsuperscript{th} of the original size (408X306). Local minima are found on the smoothed version of the image using Blom’s method as used by (Kuijper 2008).

Once the local minima are identified, the coordinate in the resized image are mapped on to the original image. This is done by multiplying the x and y coordinate values by 8. Thus on the actual image coordinate a 128 X 128 sized image region is cropped out and is stored. The image may contain region without cellular material i.e., background regions. This has to be eliminated for further analysis. For this the average intensity value of the cropped region is calculated; if the value rises above 200 it is assumed to be background since nuclear regions being darkly stained will have a low average intensity only.

The sputum cytology images are infiltrated by patches of mucus which also absorbs the stain. One more condition is given to filter out such regions. The standard deviation of the extracted region is calculated. If it is above 30 then there is an adequate distribution of dark and bright patches are present which is typical of nuclear region. All the cropped patches are saved.

The sputum sample contains a lot of artifacts like inflammatory cells, leukocytes etc which is characteristic of the sputum sampling. These artifact regions are manually removed so that the remaining dataset consists of only nuclear regions. The entire sequence of preprocessing is shown in the Figure 4.3.
Figure 4.3 Block diagram of the preprocessing system

4.6 CLUSTERING TECHNIQUES

For segmentation, two clustering techniques namely K-means and Fuzzy C-means on various color spaces such as HSV, CIELAB, CIEXYy and CIELUV are used. In the following section, details regarding the clustering algorithms, the different types of color space conversions used in the work are given. Clustering is the process of dividing the data into homogenous regions based on the similarity of objects. The process of clustering is to assign the q feature vectors into K clusters, for each k\textsuperscript{th} cluster, C\textsuperscript{k} is its center.

4.6.1 K-Means Clustering

K-means is an unsupervised clustering technique in which m data points with n dimension is grouped into k separate groups and has been pointed out by (Macqueen 1967) using the Equation (4.1).
The grouping is done by minimizing the distance between data point and the centroid of the corresponding cluster to which it falls to. \( X_j \) the data point, \( \mu_i \) the centroid of the \( i^{th} \) cluster, \( \| \cdot \|^2 \) is the l2 norm of a vector. The algorithm is run multiple times so that no more change to the data point happens. For this paper, \( k=2 \) is chosen, one for foreground and the other for background.

### 4.6.2 Fuzzy C-Means Clustering

Fuzzy clustering methods have been proposed and most of them are based upon distance criteria. The most widely used algorithm is the Fuzzy C-Mean algorithm (FCM), it uses reciprocal distance to compute fuzzy weights. This algorithm has as input a predefined number of clusters, which is the \( k \) from its name. Means stands for an average location of all the members of particular cluster and the output is a partitioning of \( k \) cluster on a set of objects. The K-means clustering assigns distinctly one group to each data point. In many applications, especially medical application, it may not be clear to which group a particular point belongs. In Fuzzy C-means a membership value is assigned to the data point for each cluster. A data point near the center of cluster has a high membership value to that particular cluster than a point near the edge of the cluster. The objective function is given using the Equation (4.2).

\[
E = \sum_{j=1}^c \sum_{i=1}^N |\mu_j|^2 \| x_i - c_j \|^2
\]  
(4.2)
\( \mu_{ij} \) is the fuzzy membership of the data point \( x_i \), \( c_j \) is the centroid, \( k \) is the constant defining fuzzyness of the cluster. After each iteration the membership function and cluster centroid is updated using the Equation (4.3) and Equation (4.4).

\[
\mu_{ij} = \frac{1}{\sum_{m=1}^{C'} \left( \frac{||x_j - c_j||}{||x_j - c_m||^{2/(k-1)}} \right)} 
\]

(4.3)

\[
C_i = \frac{\sum_{j=1}^{N} \mu_{ij}^{k} x_j}{\sum_{j=1}^{N} \mu_{ij}^{k}} 
\]

(4.4)

The FCM allows each feature vector to belong to multiple clusters with various fuzzy membership values. Then the final classification will be according to the maximum weight of the feature vector over all clusters.

### 4.7 COLOR SPACE CONVERSION

The RGB image contains pixel values which are highly correlated, so processing it directly results in redundant information processing. To uncorrelate this data, other color spaces like HSV, LUV, LAB and xyY are used. Various conversion techniques used in this work are presented below.

#### 4.7.1 RGB to HSV Conversion

The input RGB image is normalized before converting to HSV color space image using the following method was pointed out by Issac et al (2011) using the Equations (4.5) through (4.11).
\[ N = \max(R,G,B) \quad (4.5) \]

\[ n = \min(R,G,B) \quad (4.6) \]

Let, \( r = \frac{N-R}{N-n} \); \( g = \frac{N-G}{N-n} \); \( b = \frac{N-B}{N-n} \) \( (4.7) \)

Value is \( V = \max(R,G,B) \) \( (4.8) \)

Saturation is \( S = \frac{N-n}{n} \) \( (4.9) \)

Hue is
\[
H = \begin{cases} 
\frac{\pi}{3} (b - g), & \text{if } R = N \\
\frac{\pi}{3} (2 + r - g), & \text{if } G = N \\
\frac{\pi}{3} (4 + g - r), & \text{if } B = N
\end{cases} \quad (4.10)
\]

Normalized Hue is \( H = \frac{H}{2\pi} \) \( (4.11) \)

### 4.7.2 RGB to CIELAB Conversion

CIEXLAB color space is derived from CIEXYZ color space. In CIEXYZ all visible colors are represented using only positive values. The color conversion matrix is given by the Equations (4.12) through (4.15).

\[
[ R \ G \ B ] = \begin{bmatrix}
3.240479 & -1.53715 & -0.498535 \\
-0.969256 & 1.875992 & 0.041556 \\
0.055648 & -0.204043 & 1.057311
\end{bmatrix} \times \begin{bmatrix}
X \\
Y \\
Z
\end{bmatrix} \quad (4.12)
\]

The CIELAB color components are given by
\[
A = 500 \left[ h\left(\frac{X}{X_w}\right) - h\left(\frac{Y}{Y_w}\right) \right] \quad (4.13)
\]
where $X_w$, $Y_w$ and $Z_w$ are the white tristimulus values of a perfectly reflecting diffuser under CIE standard D65 illumination and have been pointed out by Gonozalez & Wood (2008).

### 4.7.3 XYZ to xyY Conversion

CIExyY is derived from CIEXYZ color space. The conversion is given using the Equations (4.16) through (4.18).

\[
x = \frac{X}{X+Y+Z} \quad (4.16)
\]

\[
y = \frac{Y}{X+Y+Z} \quad (4.17)
\]

\[Y = Y \quad (4.18)
\]

### 4.7.4 CIEXYZ to CIELUV Conversion

LUV color space is derived from XYZ color space. The conversion is given using the Equations (4.19) through (4.23).

\[
u' = \frac{4X}{X+15Y+3Z} \quad (4.19)
\]

\[
u' = \frac{9Y}{X+15Y+3Z} \quad (4.20)
\]
\[ L^* = 116 \left( \frac{Y}{Y_w} \right)^{\frac{1}{3}}, \text{ if } \frac{Y}{Y_w} > \left( \frac{6}{29} \right)^3 \]

\[ = \left( \frac{29}{6} \right)^3 \left( \frac{Y}{Y_w} \right), \text{ if } \frac{Y}{Y_w} \leq \left( \frac{6}{29} \right)^3 \]  

(4.21)

\[ u^* = 13L^*(u' - u'_w) \]  

(4.22)

\[ v^* = 13L^*(v' - v'_w) \]  

(4.23)

4.8 SEGMENTATION PROCEDURE

The algorithm used for Segmentation process is as follows:

Step 1 :  Input : 128 * 128 Image

Step 2 :  Apply K-means clustering algorithm in lab, LUV, HSV, xyY Color space.

Step 3 :  Give weight for segmented result

Step 4 :  Take the one with maximum weight

The identified nuclear regions are fed to the segmentation module. The staining of sputum cytology images generally does not give uniform color distribution. So identifying nuclear regions based on one particular color may not give good segmentation result. The information content in RGB color space is highly correlated and to uncorrelate it, various color spaces is used. Some information may be discernable in one color space may not be quiet as well distinguished in another color space.

A novel method of combining various color spaces is used in the segmentation module. The technique is a generalization of the method used
by Issac et al (2011) where K-means clustering is done on LAB color space. In this method the input image is first converted to LAB, LUV, HSV and xyY color spaces. Paired combinations of layers within these color spaces are then taken and K-mean/FCM clustering is performed upon them. The 20 combination consists of LAB layer L-A, LAB layer A-B, LAB layer L-B, HSV layer H-S, HSV layer S-V, HSV layer H-V, xyY layer x-y, xyY layer y-Y, xyY layer x-Y, LUV layer L-U and HSV layer S only. The different combinations give a different result for the segmentation using K-means and FCM algorithm.

The K-means and FCM algorithms give different labels for nuclear region in different images. So it is required to find the correct label for the nuclei region alone. The image cropping done in the preprocessing stage ensures that nuclear region is present near the middle of the image. So 100 pixels at the center of the image are taken and the median value is found. The label of the median is considered as the label for the nucleus. Then morphological closing followed by opening is performed to join small unconnected regions in the image using a 3 x 3 square structuring element. Then hole filling is done to remove holes inside nucleus region. From the resulting image the object with largest area is separated. This region gives the result of segmentation. This process is done on all the 20 results. Finally the best segmentation result has to be chosen. For this a weight value is given for each of the segmentation result. This is given by the Equation (4.24).

\[
weight = \frac{area}{convex \ area} - \frac{pixel \ in \ border}{total \ border}
\]  

(4.24)

The criterion for giving the first part is that most of the nuclear regions are of elliptical shape or shapes closer to that. So the ratio of area to convex area gives higher weight to convex shapes than concave shapes. The
K-means and Fuzzy C-means algorithm are not guaranteed to find two separate regions sometimes only one region is returned i.e., the entire image itself giving maximum weight to such an image. So this condition has to be eliminated, for this a penalty term is added.

The second term gives a penalty depending on the number of pixels that are clustered touches the boundary of the image. If no pixel is touching the border then the penalty is zero otherwise a positive value. If an erroneous clustering result comes to cover the entire image, the penalty value become one and so also the first term, giving 0 value effectively. Thus the image is eliminated from further analysis being the lowest weight value. From the twenty segmented results one with maximum weight is taken as the final segmented result. The various step of segmentation are shown in Figure 4.4.

Figure 4.4 The steps of segmentation algorithm
4.9 FEATURE EXTRACTION USING SCALE SPACE

Scale space deals with the observation or study of image at various scales. It is about describing an image according to the scale of objects present in the image. In this paper, scale space catastrophe points as features are used. Catastrophe points or top points were successfully used for image matching and reconstruction which was addressed by Kimmel et al (2011). In this work, catastrophe points are used for extracting features from the nuclear region.

In the physical world every object has the corresponding property which is observed occur at certain scale of observation only. That is to know about an object that is analyze occur at particular scale which is meaningful to the under study. This forms the basis of scale space theory and was discussed by (Lindeberg 2000). Convolving the image with a Gaussian kernel gives the linear scale space representation and has been pointed out by (Koenderink 1984). Linear scale space was the unique kernel which satisfies the following axioms and was given by Florack et al (1994). linearity, spatial scale invariance, spatial homogeneity and spatial isotropy

It was shown that if separability is not a criterion then there exist infinite linear scale spaces and has been pointed out by Duits et al (2004). In this analysis the linear Gaussian scale space is considered.

The scale space stack is generated and stored for further analysis. There are certain interesting events which happen in the scale space. This can be analyzed only by observing the scale space stack as a whole i.e. both along spatial and scale dimensions and this is termed as deep structure (Koenderink 1984). In general deep structure analysis shows that the singularities decrease as the scale space stack is moved up and was given by (Romemy 2003). This
is due to the annihilation and creation of singularities depending on the scale of observation.

Observing singularities alone in the scale space stack gives a basic idea of the image structure. Analyzing the singularity leads to the catastrophe points where major change to scale space image structure occurs and was pointed out by Florack et al (1994).

Maxima, minima and saddle points are found on all the scales using Blom’s method. Here hexagonal neighborhoods are considered for processing. Even rows in rectangular grid are shifted half pixel to the right to form the hexagonal grid. In practice this turns out to be taking the eight nearest neighbors and discarding top right and bottom right neighbors for even rows and discarding top left and bottom left elements for odd rows (Kuijper 2004). If no sign change, the given point is extrema and if two sign changes, the point is a regular point. If four sign changes, the points are referred as a saddle and if six sign changes, the point is a money saddle. The Figure 4.5 shows how the catastrophe points are found on all the scales.

**Figure 4.5 The catastrophe points on different scale**

The maxima, minima and saddle points found are arranged over all the scales. At each level these critical points are connected to the level above it using adjoining neighbor approach. The linking is done at all the scales.
Whenever the critical points reach their annihilation catastrophe they cease to exist in the succeeding higher scales. The level at which this happens is identified and marked as the catastrophe level of that critical point. The saddle-maxima and saddle minima catastrophe is found using nearest neighbor linking.

The experimental results show that the number of catastrophe points is more for malignant cases than for benign cases. Scale space stack is generated for 32 scales ranging from $e^1$ to $e^3$. The intermediate scales are chosen in such a way that the ratio between any two successive scales is the same and has been pointed out by (Lindeberg 1998). The count for catastrophe points, both saddle-maxima and saddle minima, occurring for each scale is taken and the feature vector is generated thus producing a size of 64 dimensional vectors.

### 4.10 CLASSIFICATION

The classifier forms the final part of the system. In this classification stage, the individual cells are classified as benign or malignant. The cell classification is done in two stages. In stage one, the nuclei are classified as benign or malignant using SVM and in stage two, those nuclei which are classified as malignant is analyzed along with cytoplasm using Z-test.

SVM allows for a training error in which to maximize the margin some misclassification is allowed which results in better classification during testing. For high dimensional data, linear separation may not be possible to achieve. To facilitate nonlinear mapping, kernel trick is applied in which the input data point is mapped to a higher dimension using dot product of input features.
After the classification of cell nuclei as benign or malignant by SVM, the cells classified as malignant is rechecked by analyzing the cytoplasm, for which Z-test is applied and the entire process is shown in the form of a flow chart in the Figure 4.6. Here the stain absorption of cytoplasm is assumed to be Gaussian. Z-test is a statistical hypothesis testing in which a null hypothesis is tested against an alternative hypothesis. The test is done to analyze whether the average color intensity is that of a benign cell. The null hypothesis is that the average color intensity of the cytoplasm is given by \{174, 166, 154\} with standard deviation \{8.4, 9, 9\} for red, green and blue channels respectively for a population size of 1000 cell images. Thus there are altogether three two tailed tests one for each color channel done at 5% significance level. If the entire three tests pass the null hypothesis then the cytoplasm is considered to be that for benign and those cells are categorized as benign with reactive changes. All the remaining cells form malignant category.

Figure 4.6 Flow chart to give the stages of classification process
The training set consists of 100 cells from both benign and malignant cases. The classification of the selected cells was done manually by expert cytopathologist. The testing sample consists of 768 benign and 270 malignant cells. In this work, the non-linear Radial Basis Function (RBF) kernel for SVM is used. The accuracy of the classifier is as shown in Table 4.1.

Table 4.1 Accuracy using SVM RBF Kernel

<table>
<thead>
<tr>
<th>Performance Parameter</th>
<th>No. of cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>112</td>
</tr>
<tr>
<td>True Negative</td>
<td>704</td>
</tr>
<tr>
<td>False Positive</td>
<td>64</td>
</tr>
<tr>
<td>False Negative</td>
<td>158</td>
</tr>
<tr>
<td>Accuracy %</td>
<td>78.61</td>
</tr>
</tbody>
</table>

4.11 RESULTS AND DISCUSSION

The cellular region from input high resolution image 3264 X 2448 is cropped out and is shown in Figure 4.7

![Figure 4.7 Cropped nucleus from the input image](image)

Majority of the cellular regions were detected and fed to the subsequent segmentation stage. In this stage the exact region of nucleus is
delineated. Cropped cellular regions for benign and malignant cells are shown in Figure 4.8. Figure 4.8 a) shows a sample cropped benign cell. Figure 4.8 b) shows a sample cropped malignant cell.

![Figure 4.8 Cropped benign and malignant cell](image)

The cropped cellular region is subsequently processed in the segmentation module where the exact nuclear region is obtained. For segmentation of nuclear region, 20 results are obtained using various combinations of color spaces and clustering using FCM and K-means from which the best is chosen. The segmentation result along with clustering technique, color space, color channels used and corresponding weight is shown in Figure 4.9 for malignant cell.

![Figure 4.9 Segmentation results for malignant cell using color spaces](image)
Figure 4.10 shows the segmentation result along with clustering technique, color space, color channels used and corresponding weight of a sample benign cell for the given sputum cytology image. The best segmentation result follows to the feature extraction stage where the catastrophe points are extracted.

Figure 4.10 Segmentation results for benign cell using color spaces

Maxima (represented as part a in the first plot), Minima (represented as part b in the second plot) and Saddle (represented as part c in the third plot) are detected from the images in the scale space stack and are shown in Figure 4.11.
Figure 4.11 Maxima, Minima and Saddle for various scales

Experimental results for benign case 1 and benign case 2 with catastrophe points on different scales are shown in the Figure 4.12 and Figure 4.13 respectively.

Figure 4.12 Experimental result for benign case 1

Figure 4.13 Experimental result for benign case 2
Figure 4.13 Experimental result for benign case 2

Experimental results for malignant case 1 and malignant case 2 with catastrophe points on different scales are shown in the Figure 4.14 and Figure 4.15 respectively.

Figure 4.14 Experimental result for malignant case 1
It is observed from the Figure 4.12 through Figure 4.15 that the numbers of catastrophe points are more for malignant cases than the benign cases. This helps us to classify benign and malignant cells from the lung glandular cells. A Sample malignant cell and its corresponding Maxima, minima and Saddle points are shown in the Figure 4.16.

Figure 4.16 Sample malignant cell and its corresponding connected catastrophe points
A sample benign cell and its corresponding Maxima, minima and Saddle points are shown in the Figure 4.17. It is observed that the distributions of critical points for both malignant and benign cells are visually different. Presence of more maxima points is clear in the case of malignant cells as compared to benign cells.

![benign cell 1](image)

**Figure 4.17 Sample benign cell and its corresponding connected catastrophe points**

### 4.12 SUMMARY

In this work, a novel semi automated method for lung glandular cell segmentation using multiple color spaces and two clustering techniques namely FCM and C-Means are given. Catastrophe points are used as features and SVM is used as classifier for the purpose. It was observed that the circulation of critical points for both malignant and benign cells are visually dissimilar. Occurrence of extra maxima points was clear in the case of malignant cells as compared to benign ones. In this work, the nuclear region in terms of maxima, minima and saddle in the scale space stack are
represented. In this method, local maxima in saddle and extrema points and take it as interest point for future modeling of the nucleus was found. The initial experiments show that there is a definite visual difference in the appearance of benign and malignant nuclei. The system successfully identifies the cellular regions as benign or malignant when the cropped region is fed with an accuracy of 78.61%.

As future work, it is to improve the accuracy by adding more features and a fully automated system which will automatically remove artifacts and compare the results with other features and use different classification techniques.