Creating a Optimal Set of Textural Features for Cervical Cancer Lesions Using Hierarchal Clustering Technique

S. Pradeep Kumar Kenny¹, Dr. S. Allwin²

¹Centre for Information Technology & Engineering, Manonmaniam Sundaranar University, Tirunelveli
²Dept. of Computer Science and Engineering, Infant Jesus College of Engineering and Technology, Tuticorin, India.

Abstract—Cervical cancer is considered to be the deadliest gynaecological cancers which affect women and it now seems that the only way it could be controlled is by employing an automated system. Of all the visual parameters which the brain uses to make sense of the world around us the textural features is the best reliable one. Here we have tried to exploit this concept by analyzing the textural features of cervical cyto images. The Nucleus of a cervical cyto image has with it varied textural information which inherently tell about the stage of the cancer. There are many textural features available and since each image has different properties no two images of the same stages will show positive for a particular feature. Hence the alternative is to have combination of features. But in doing so you must have a right balance between speed and accuracy of the system. Hence in this work apart from extracting the features we have proposed a ranking system which will help us to create a balanced set of combinational features so that it won’t affect the accuracy nor the speed of the automated system.

Keywords—Cancer, Cervical Cancer, Cervical Cytology, Clustering, Hieratcal Clustering, Textural Features,

I. INTRODUCTION

Cervical cancer is the cancer occurring in the cervix of the women’s reproductive system. The cervix is also called the birth canal and it connects the uterus to the vagina. Cervical cancer is the second most common cancer among women and it also causes many deaths among women worldwide. The reason for this is wide ranged from lack of proper medical facilities to unawareness of the disease etc.,. The only way to prevent this cancer is to undergo periodic screening. The most commonly used screening technique is called the Pap Smear test. In which, a tissue sample is taken from the cervix using a cotton swab or a stick by swabbing the region. Then the cells are analysed under a microscope by a doctor who with his knowledge tells whether the particular patient has cancer or not. This approach has a lot of problems not only because it is time consuming also it is not possible for a handful of doctors to do the screening for billions of women worldwide. An alternate to this would be an automated system which could analyse and tell the possibilities of cancer and thus reduce the workload of the doctor and increase the amount of women been screened.

Ever since the computers were invented mankind has been looking for ways to make it do all our work. Computers are known immensely for their valuable characteristics and when we speak of an automated system we can’t think anything more than a computer. But before we can use a computer as a automated system a thorough analysis has to be performed as to which features extracted would make the system work better. There are many automated system now being designed but the crucial factor for all is that they should work with accuracy and speed. In these kinds of systems they have to be balanced. Increasing or decreasing one other would result in dire consequences. So the focal point is to focus on the features used as that is one which clearly falters the automated system.

An image is described has what it is by the features it processes. It can be used to describe what the image actually is. Here we have taken into consideration four different classes of textural features. Textural features yield important information about an object. Any cancer cell, progresses through a series of stages before it becomes invasive. The mitosis process governs cell division and growth of various parts in the human body. Each part of the human body stops growing at some stage which implies that the mitosis process should also stop. But in some cases the mitosis process continues even though the growth of a certain body part has stopped. This results in a situation where the divided cells don’t have enough space to occupy and the cells tend to occupy the cytoplasm region. Once the cells find no more room they clump over one another causing tumours to form. This is the onset of the precancerous stage where cancers are normally detected. The accumulation of cells continues and forms the stage where the cancer has spread to neighbouring regions and is very difficult to cure beyond this point.

There are many textural features available but for a system to work successfully we need the right number of features to be extracted. In this paper we have taken four classes of textural features and by employing a unique classification technique we intend to rank the cells. This paper has been organized into IV sections. Section II discusses how the features extracted and ranked. And Section III shows the results. Conclusions are made in Section IV.
II. Analysis For Textural Features In Cervical Cyto Images

The proposed method is an effective method that can be applied on cervical cyto images. The Block Diagram for this approach is given in Figure 1.

<table>
<thead>
<tr>
<th>Input</th>
<th>Pre-processing</th>
<th>Feature Extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A. Input Image

The input image is the tissue of the cervix being digitalized. A sample of a cervical cell is shown in figure 2[1]. Here in this paper we have used about 917 images which are pre classified images obtained from the Herlev dataset. Cyto technicians and doctors have manually classified each cell into one of the 7 classes, namely superficial squamous, intermediate squamous, columnar, mild dysplasia, moderate dysplasia, severe dysplasia and carcinoma in situ. The first three correspond to normal classes and the last four classes correspond to abnormal classes as shown in figure 3[1]. The pre-classification helps to serve as the gold standard for the entire work[2].

B. Pre-Processing

The input images are represented by a 3 dimensional color model which increases computational time if processed. Hence the 3 dimensional color model is decomposed to a two dimensional color model. Also the Region of interest need to be isolated from the rest of the image by a process called segmentation. There are numerous methods proposed for isolating these regions of interest [3]-[13]. Anyone of these methods can be used for isolating the regions of interest.

The pre-processing technique used in this paper first converts the image to 2 dimensional colour model and then filters the noise with the median filter and then uses a set of morphological operators to strengthen the image. At the end we end up getting an image as shown in figure 4[14]. The equation for obtaining this is shown below.

\[
C(f)_p = \frac{P(f)_p}{\sum_{i=1}^{M}(P(f)_p+M(f)_p)}; \quad p \text{ is black in } M(f)
\]

C. Feature Extraction

Texture composition is the one that yields a wide spectrum of information about a certain image. Our brain uses this feature to describe what an image is. There are several textural features available but in this paper we are taking into consideration the gray level co-occurrence matrix (GLCM) features, Haralick, gradient and tamura based features.

<table>
<thead>
<tr>
<th>Normal Cells</th>
<th>Abnormal Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superficial Squamous</td>
<td>Mild Dysplasia</td>
</tr>
<tr>
<td>Intermediate Squamous</td>
<td></td>
</tr>
<tr>
<td>Columnar</td>
<td></td>
</tr>
</tbody>
</table>

Moderate Dysplasia

Severe Dysplasia

Carcinoma in Situ

Figure 3. Sample images of the Herlev Dataset

Figure 2. Cervical Cyto Image

Figure 3 Segmented Image

GLCM is one of the simplest ways used to describe the features in a given image. It makes use of statistical moments of the intensity histogram of an image and is described mathematically as follows.
\[ P(i, j) = \sum_{z=1}^{N} \sum_{y=1}^{N} \left( \begin{array}{cl} 1, & \text{if } f(x, y) = \text{is and }l(x+\Delta_x, y+\Delta_y) = j \\ 0, & \text{otherwise} \end{array} \right) \]  

(2)

Where \( i \) and \( j \) are the image intensity values, \( p \) and \( q \) are the spatial positions in an image \( l \) and the offset \( (\Delta_x, \Delta_y) \) specifies the distance between the pixel-of-interest and its neighbour.

In 1973 Haralick introduced 14 statistical features which are generated by calculating the features for each one of the co-occurrence matrix obtained by using the directions \( 0^\circ, 45^\circ, 90^\circ \) and \( 135^\circ \). Then by averaging these four values. The Symbol \( \Delta \) representing the distance parameter, can be selected as one or higher. In general \( \Delta \) can be set to 1 as the distance parameter. Where \( E \in \mathbb{Z}^2 \times \mathbb{R} \) is the extinction image after intensity conversion.

A 3x3 gradient approximation is used. This is an approximation of \( \left( \frac{\partial I(x,y)}{\partial x}, \frac{\partial I(x,y)}{\partial y} \right) \). The \( G \ldots \) features can be considered as a quantification of the velocity of gray values. The measuring field is eroded by a 3x3 square[15]. These values are then plotted using the following equations.

\[ h = \text{histogram} (X(O)) \]  

(3)

\[ h(v) = \text{frequency of pixel value of } v \]  

(4)

Tamura Image is a notion where we calculate a value for the three features at each pixel and treat these as a spatial joint coarseness-contrast-directionality (CND) distribution.

In this paper we have employed around 34 features for our analysis. The features implemented are listed below [15]-[19].

1) \( \text{Mean:} \)

\[ M1 = \frac{1}{M0} \sum_v v^4 h(v) = \sum_v v H(v) \]  

(5)

2) \( \text{Standard deviation:} \)

\[ M2 = \frac{1}{M0} \sum_v v^2 h(v) - (M1)^2 \]  

(6)

3) \( \text{Median:} \)

\[ MED = \text{median} (X(O)) \]  

(7)

4) \( \text{Entropy:} \)

\[ \text{ENT} = - \sum_v H(v) \log_2(H(v)) \text{ entropy of } h \]  

(8)

5) \( \text{Nucleocytoplasmic Ratio (NCR):} \)

\[ NCR = \frac{Na}{Ca} \]  

(9)

6) \( \text{Sum Variance:} \)

\[ f_7 = \sum_{i=2}^{2N} (i - f_8)^2 P_{x,y}(i) \]  

(10)

7) \( \text{Contrast:} \)

\[ \sum_{i,j=0}^{N-1} P_{i,j} (i-j)^2 \]  

(11)

8) \( \text{Hyperchromasia} \)

\[ T_w(f) = f - f \circ b \]  

(12)

9) \( \text{Sum} \)

\[ M0 = |O| \]  

(13)

10) \( \text{Skewness} \)

\[ \frac{1}{(M2)^3} \left[ \sum v^3 h(v) - 3 \sum v^2 h(v) \sum v h(v) + 2 \sum v^8 h(v) \right] \]  

(14)

11) \( \text{Kurtosis} \)

\[ \frac{1}{(M2)^4} \left[ \sum v^4 h(v) - 4 \sum v^3 h(v) \sum v^2 h(v) - 3 \sum v^6 h(v) \right] \]  

(15)

12) \( \text{Max} \)

\[ \max(X(O)) \]  

(16)

class of largest non-zero frequency of \( h \)

13) \( \text{Autocorrelation} \)

\[ f_6 = \sum_i \sum_j (i,j)(p(i,j)) \]  

(17)

14) \( \text{Correlation} \)

\[ \sum_{i,j=0}^{N-1} P_{i,j} \left[ \frac{(i-\mu_i)(j-\mu_j)}{(\sigma_i^2)(\sigma_j^2)} \right] \]  

(18)

15) \( \text{Cluster Prominence} \)

\[ f_8 = \sum_i \sum_j (i+j-\mu_x+\mu_y)^p p(i,j) \]  

(19)

16) \( \text{Cluster shade} \)

\[ f_7 = \sum_i \sum_j (i+j-\mu_x+\mu_y)^3 p(i,j) \]  

(20)

17) \( \text{Dissimilarity} \)

\[ f_6 = \sum_i \sum_j |i-j| p(i,j) \]  

(21)
18) Energy

\[ f_1 = \sum_{i} \sum_{j} \{p(i, j)\}^2 \]  

(22)

19) Entropy of GLCM

\[ f_9 = -\sum_{i} \sum_{j} p(i, j) \log (p(i, j)) \]  

(23)

20) Homogeneity

\[ f_5 = \sum_{i} \sum_{j} \frac{1}{1+(i-j)^2} p(i, j) \]  

(24)

21) Maximum Probability

\[ f_{10} = \underset{i, j}{\text{MAX}} p(i, j). \]  

(25)

22) Sum of average

\[ f_6 = \sum_{i=2}^{2N-1} i P_{x+y}(i) \]  

(26)

23) Sum of square variance

\[ \sigma_i^2 = \sum_{i,j=0}^{N-1} P_{i,j} (i-\mu_i)^2, \sigma_j^2 = \sum_{i,j=0}^{N-1} P_{i,j} (j-\mu_j)^2 \]  

(27)

24) Sum of entropy

\[ f_8 = \sum_{i=2}^{2N-1} P_{x+y}(i) \log \{P_{x+y}(i)\} \]  

(28)

25) Difference variance

\[ f_{10} = \text{variance of } P_{x-y} \]  

(29)

26) Difference Entropy

\[ f_{11} = -\sum_{i=0}^{N-1} P_{x-y}(i) \log \{P_{x-y}(i)\} \]  

(30)

27) Information Measurement of Correlation 1 & 2

\[ f_{12} = \frac{HXY - HXY_1}{\text{MAX} \{HXY, HX, HY\}} \]  

(31)

\[ f_{13} = (1 - \exp [-2.0(HXY_2 - HXY)])^{1/2} \]  

(32)

\[ HXY = -\sum_{i} \sum_{j} p(i, j) \log (p(i, j)) \]  

(33)

where HX and HY are entropies of p_x and p_y, and

\[ HXY_1 = -\sum_{i} \sum_{j} p(i, j) \log (p_x(i)p_y(j)) \]  

(34)

\[ HXY_2 = -\sum_{i} \sum_{j} p_x(i)p_y(j) \log (p_x(i)p_y(j)) \]  

(35)

28) Inverse difference

\[ \sum_{i,j} \frac{C_{i,j}}{1 + |i - j|^2} \]  

(36)

29) Inverse difference normalized

\[ \sum_{i,j} \frac{C_{i,j}}{1 + |i - j|^2 / G^2} \]  

(37)

30) Inverse difference moment normalized

\[ \sum_{i,j} \frac{C_{i,j}}{1 + (i - j)^2 / G^2} \]  

(38)

31) Coarseness

It involves implementation of the following steps

Step 1:-

For every point \( n_0, n_1 \) calculate differences between the not overlapping neighborhoods on opposite sides of the point in horizontal and vertical direction:

\[ E_1^2(n_0, n_1) = |A_4(n_0 + 2^{k-1}, n_1) - A_4(n_0, n_1 - 2^{k-1})| \]  

(39)

Step 2:-

At each point \( n_0, n_1 \) select the size leading to the highest difference value:

\[ S(n_0, n_1) = \arg\max_{k=1, \ldots, 5} \max_{d=1} E_k^d(n_0, n_1) \]  

(40)

Step 3:-

Finally take the average over 2S as a coarseness measure for the image:

\[ F_{crs} = \frac{1}{N_0N_1} \sum_{n_0=1}^{N_0} \sum_{n_1=1}^{N_1} S(n_0, n_1) \]  

(41)

32) Directionality

\[ \theta = \frac{\pi}{2} + \tan^{-1} \frac{\Delta v(n_0, n_1)}{\Delta v(n_0, n_1)} \]  

(42)

33) Tamura feature Contrast

\[ F_{con} = \frac{\sigma^2}{\sigma^4} \text{ with } \sigma_4 = \frac{\mu_4}{\sigma^4} \]  

(43)
Where

- \( N_a \) is the number of pixels in the cell nucleolus region
- \( C_a \) is the number of pixels in the cytoplasmic region
- \( f \) is the grayscale image
- \( b \) is the structuring element
- \( \circ \) is the opening operation
- \( O \) is the total number of elements in the ROI
- \( h(v) \) frequency of pixel value \( v \)
- \( M0 \) is the total number of elements in the ROI
- \( M1 \) is the mean
- \( M2 \) is the standard deviation
- \( X(O) \) is the set of all pixels of \( X \) in \( O \) which is the class of largest non – zero frequency of \( h \)
- \( p(i,j) \) is the \( i,j^{th} \) entry in the normalized graytone spatial dependence matrix
- \( P(i,j)/R \) and \( N_g \) is the number of distinct gray levels in the quantized image.
- \( \mu_x, \mu_y \) \( \sigma_x \) and \( \sigma_y \) are the means and standard deviations of \( p_x \) and \( p_y \).
- \( \mu_x = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \) and \( N_g \) is the number of distinct gray levels in the quantized image.
- \( \sum \) is the Sum of Entropy features

\[ \text{HX and HY are entropies of } p_x \text{ and } p_y \]
\[ \mu_x = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j), \mu_y = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \]
\[ C_{ij} = \frac{p_{ij}}{\sum_{j=1}^{N_g} p_{ij}} \]
\[ \mu_4 = \frac{1}{N_g^2} \sum_{n=1}^{N_g} \sum_{m=1}^{N_g} (X(n, m) - \mu)^4 \]
\[ \sigma^2 \text{ is the variance of the gray values present in the image and } z \text{ has to be experimentally determined to be } \frac{1}{4} \]

**D. Outlier Deletion**

The extracted features are plotted in a graph to help in removing the outliers. Outliers are the features that won’t help in detecting a particular stage of cancer. Each feature is extracted for all the images and they are plotted as shown in Figure 4.

To help perform this grubs test is being used which is based on the difference of mean and the most extreme data considering the standard deviation given dataset with assumed normal distribution. The test is based on the following formulas

\[ T_{\text{max}} = \frac{X_n - X_{\text{mean}}}{S} \]
\[ T_{\text{min}} = \frac{X_{\text{mean}} - X_j}{S} \]

Where

- \( X_n \) or \( X_j = \) the suspected single outlier (max or min)
- \( S \) = standard deviation of the whole data set
- \( X_{\text{average}} = \text{mean} \)

Once the outliers are deleted the total percentage of images present are calculated using the following formula

\[ \% \text{ of correct detection} = \frac{\Delta o}{\Delta s} \]

Where

- \( \Delta o \) is the total images present in a stage after deletion of outliers
- \( \Delta s \) is the total images present in a the given dataset for a particular stage

**E. Ranking**

Once outlier deletion is performed we are left with values or in other words features that work best for an image. Now we can ofcourse use a threshold and say values above a said threshold can be used but it won’t guarantee the accuracy of the system. Hence a unique ranking system is used. The stages are arranged column wise and features are arranged row wise. Each feature has higher predominance for a particular stage of cancer. So if you look from the row perspective you can find it out clearly. A clustering scheme can help to cluster the values of high predominance. A hierarical clustering is employed here to do just that and it makes use of a dendrogram to do it.

A sample dendrogram for the autocorrelation GLCM textural feature is shown in figure 5. The x axis represents the stages. As you can see there is union between the 1 and 2 stage in the first iteration, union between stage 3 and 6 in the second iteration, union between stage 4 and 5 in the third iteration. All these iteration output combined to form the union in iteration four and a union with this result and stage 7 completes the clustering part. Now the order of the union passing through each iterations is the rank. Therefore it can be said that the prominence for autocorrelation is stage 1 and 2 and henceforth.
One is to see in terms of features. For eg:- you take NCR features and it ranks first for the carcinoma in situ stage. But if you look in terms of carcinoma in situ column you have certain features with rank 1’s. These can form the combinational set of features. The entire execution was performed using MATLAB Version 7.10.0.499 and since it reads alphabetical from folders while inputting images, order of the table has been changed.

### TABLE I

Percentage of liable images after outlier deletion

<table>
<thead>
<tr>
<th>F</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.54</td>
<td>75.71</td>
<td>84.69</td>
<td>86.26</td>
<td>82.88</td>
<td>88.83</td>
<td>74.67</td>
</tr>
<tr>
<td>2</td>
<td>82.43</td>
<td>78.57</td>
<td>84.69</td>
<td>73.08</td>
<td>87.67</td>
<td>81.73</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>85.14</td>
<td>84.29</td>
<td>83.67</td>
<td>85.71</td>
<td>81.51</td>
<td>86.8</td>
<td>83.33</td>
</tr>
<tr>
<td>4</td>
<td>74.32</td>
<td>74.29</td>
<td>79.59</td>
<td>78.02</td>
<td>77.4</td>
<td>88.83</td>
<td>83.33</td>
</tr>
<tr>
<td>5</td>
<td>81.08</td>
<td>72.86</td>
<td>87.76</td>
<td>90.66</td>
<td>84.25</td>
<td>84.77</td>
<td>84</td>
</tr>
<tr>
<td>6</td>
<td>81.08</td>
<td>72.86</td>
<td>81.63</td>
<td>89.01</td>
<td>83.56</td>
<td>87.82</td>
<td>83.33</td>
</tr>
<tr>
<td>7</td>
<td>79.73</td>
<td>72.86</td>
<td>74.49</td>
<td>87.91</td>
<td>74.66</td>
<td>83.25</td>
<td>79.33</td>
</tr>
<tr>
<td>8</td>
<td>83.78</td>
<td>72.86</td>
<td>87.76</td>
<td>84.62</td>
<td>87.67</td>
<td>82.74</td>
<td>91.33</td>
</tr>
<tr>
<td>9</td>
<td>81.08</td>
<td>87.14</td>
<td>89.8</td>
<td>88.46</td>
<td>78.77</td>
<td>79.7</td>
<td>84.67</td>
</tr>
<tr>
<td>10</td>
<td>87.84</td>
<td>50</td>
<td>84.69</td>
<td>53.3</td>
<td>87.67</td>
<td>85.79</td>
<td>93.33</td>
</tr>
<tr>
<td>11</td>
<td>72.97</td>
<td>88.57</td>
<td>96.94</td>
<td>74.18</td>
<td>85.62</td>
<td>82.23</td>
<td>82</td>
</tr>
<tr>
<td>12</td>
<td>85.14</td>
<td>81.43</td>
<td>81.63</td>
<td>84.07</td>
<td>85.62</td>
<td>87.82</td>
<td>97.33</td>
</tr>
<tr>
<td>13</td>
<td>89.19</td>
<td>80</td>
<td>77.35</td>
<td>91.21</td>
<td>87.67</td>
<td>77.16</td>
<td>92.67</td>
</tr>
<tr>
<td>14</td>
<td>90.54</td>
<td>92.86</td>
<td>87.76</td>
<td>79.67</td>
<td>88.36</td>
<td>90.36</td>
<td>85.33</td>
</tr>
<tr>
<td>15</td>
<td>87.84</td>
<td>84.29</td>
<td>82.65</td>
<td>89.56</td>
<td>83.56</td>
<td>71.07</td>
<td>74.67</td>
</tr>
<tr>
<td>16</td>
<td>91.89</td>
<td>92.86</td>
<td>80.61</td>
<td>82.42</td>
<td>87.67</td>
<td>86.8</td>
<td>86.67</td>
</tr>
<tr>
<td>17</td>
<td>90.54</td>
<td>92.86</td>
<td>87.76</td>
<td>79.67</td>
<td>88.36</td>
<td>90.36</td>
<td>85.33</td>
</tr>
<tr>
<td>18</td>
<td>72.97</td>
<td>77.14</td>
<td>85.71</td>
<td>78.02</td>
<td>85.62</td>
<td>79.19</td>
<td>82</td>
</tr>
<tr>
<td>19</td>
<td>85.14</td>
<td>84.29</td>
<td>80.61</td>
<td>89.01</td>
<td>74.66</td>
<td>80.71</td>
<td>85.33</td>
</tr>
<tr>
<td>20</td>
<td>86.49</td>
<td>87.14</td>
<td>80.61</td>
<td>86.81</td>
<td>85.62</td>
<td>82.74</td>
<td>78</td>
</tr>
<tr>
<td>21</td>
<td>85.14</td>
<td>82.86</td>
<td>79.59</td>
<td>86.26</td>
<td>78.77</td>
<td>77.66</td>
<td>72</td>
</tr>
<tr>
<td>22</td>
<td>72.97</td>
<td>91.43</td>
<td>83.67</td>
<td>80.22</td>
<td>89.73</td>
<td>86.29</td>
<td>80</td>
</tr>
<tr>
<td>23</td>
<td>81.08</td>
<td>94.29</td>
<td>83.67</td>
<td>71.98</td>
<td>87.67</td>
<td>82.74</td>
<td>81.33</td>
</tr>
<tr>
<td>24</td>
<td>91.89</td>
<td>94.29</td>
<td>80.61</td>
<td>82.42</td>
<td>82.19</td>
<td>85.79</td>
<td>80.67</td>
</tr>
<tr>
<td>25</td>
<td>75.68</td>
<td>82.86</td>
<td>80.61</td>
<td>66.48</td>
<td>89.73</td>
<td>79.7</td>
<td>86.67</td>
</tr>
<tr>
<td>26</td>
<td>90.54</td>
<td>90</td>
<td>83.67</td>
<td>85.71</td>
<td>79.45</td>
<td>77.66</td>
<td>84.67</td>
</tr>
<tr>
<td>27</td>
<td>71.62</td>
<td>85.71</td>
<td>85.71</td>
<td>87.36</td>
<td>87.67</td>
<td>77.16</td>
<td>75.33</td>
</tr>
<tr>
<td>28</td>
<td>85.14</td>
<td>87.14</td>
<td>80.61</td>
<td>81.32</td>
<td>82.19</td>
<td>83.25</td>
<td>78.67</td>
</tr>
<tr>
<td>29</td>
<td>83.78</td>
<td>87.14</td>
<td>75.51</td>
<td>90.11</td>
<td>75.34</td>
<td>81.22</td>
<td>77.33</td>
</tr>
<tr>
<td>30</td>
<td>77.03</td>
<td>82.86</td>
<td>87.76</td>
<td>88.46</td>
<td>82.19</td>
<td>86.8</td>
<td>90</td>
</tr>
<tr>
<td>31</td>
<td>79.73</td>
<td>77.14</td>
<td>89.8</td>
<td>59.34</td>
<td>78.77</td>
<td>80.71</td>
<td>85.33</td>
</tr>
<tr>
<td>32</td>
<td>87.84</td>
<td>90</td>
<td>79.59</td>
<td>75.27</td>
<td>84.93</td>
<td>76.14</td>
<td>80.67</td>
</tr>
<tr>
<td>33</td>
<td>72.97</td>
<td>81.43</td>
<td>88.65</td>
<td>78.02</td>
<td>84.25</td>
<td>88.32</td>
<td>92</td>
</tr>
<tr>
<td>34</td>
<td>74.32</td>
<td>71.43</td>
<td>82.65</td>
<td>78.57</td>
<td>83.56</td>
<td>84.77</td>
<td>85.33</td>
</tr>
</tbody>
</table>

Where

<table>
<thead>
<tr>
<th>Abnormal Cells</th>
<th>Normal Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: Carcinoma in Situ</td>
<td>S5: Columnar</td>
</tr>
<tr>
<td>S2: Mild Dysplasia</td>
<td>S6: Intermediate Squamous</td>
</tr>
<tr>
<td>S3: Moderate Dysplasia</td>
<td>S7: Superficial Squamous</td>
</tr>
<tr>
<td>S4: Severe Dysplasia</td>
<td>F: Features</td>
</tr>
</tbody>
</table>

**Note:** The order of the stages has been changed to help in computation.
work. In future more feature set can be employed to predefined classified dataset helped the accuracy of this each stage of cancer. Performing these result on top of a Tamura feature Contrast Directionality Coarseness

TAMURA FEATURES

Inverse difference normalized 3 7 4 6 1 5 2
Inverse difference moment normalized 3 7 4 5 1 2 6

Note: Order of features are changed for representation.

IV. CONCLUSION

Thus we have created a refined set of feature set for each stage of cancer. Performing these result on top of a predefined classified dataset helped the accuracy of this work. In future more feature set can be employed to strike a more versatile combination of features.

REFERENCES

[17] Prof. Dr.-Ing.H.Ney., Features for Image retrieval, December 2003
Figure 6 Dendrogram for all the features.

Note: The order of features are same as that of Table I