ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(19), DOI: 10.17485/ijst/2016/v9i19/93881, May 2016* ISSN (Online) : 0974-5645

Feature Extraction and Feature Set Selection for Cervical Cancer Diagnosis

G. Karthigai Lakshmi[1*] and K. Krishnaveni[2]

[1]Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India; karthigailakshmi64@gmail.com
[2]Department of Computer Science, Sri Ramasamy Naidu Memorial College, Sattur, Tamil Nadu, India; kkveni_srnmc@yahoo.co.in

Abstract

Objectives: To identify a cervical cytology cell image as a representative of the normal or malignant cancerous cell. This work may be used to decide the stage of cervical cancer. **Methods and Analysis:** Images from the Herlev university database have been used. Features of nucleus and cytoplasm of cervical cell images are extracted after preprocessing and segmentation. Size, shape and texture features are extracted. Appropriate and adequate features are selected by mining techniques. This feature set is fed to classifiers to decide nature of cells. Based on parameters like true positive rate, true negative rate and precision, a comparative analysis of classifiers is done. The best classifier that clearly discriminates the cells is determined for cervical cancer diagnosis. **Findings:** Dimensionality reduction of feature set can be done with data mining technique named as Correlation based feature subset selection. Reduced feature set can be used for classification by classifiers like Multilayer perceptron, Bayes classifiers and SVM classifier. Kappa Statistics for binary classifiers is 1 for Multilayer perceptron, 0.9968 for Binary SMO and Bayesnet, 0.9915 for NaiveBayes Classifiers. The same for seven group classifiers is 1 for LibSVM, 0.8535 for Multilayer perceptron, 0.8464 for attribute selected classifier and 0.7286 for Bagging classifier. True positive rate is 1 for Multilayer perceptron, 0.996 for Binary SMO and BayesNet, 0.975 for NaiveBayes classifier. Multilayer perceptron can be used as a classifier in both binary and seven class classifications. **Improvement:** Existing methods use more features for classification. Here even in the case of seven class classifier, only 11 features are used and desired result has been achieved.

Keywords: Classifiers, Clustering, Data Mining, Dimensionality Reduction, Image Texture Analysis

1. Introduction

Cervical cancer is one of the deadly cancers affecting the mortality rate of women in developing countries. In United States of America, estimated new cases in the year 2015 is 12,900[1] and estimated all new cases percentage is only 0.8% since annual screening of cervical cancer is mandatory in United States of America. But in India, according to statistics of 2015[2], every year 1,22,844 women are diagnosed with cervical cancer and 67,477 die from the disease. This cancer ranks as the second most frequent cancer affecting women between 15 and 44 years of age, in India. Since the mortality rate of women affected by this disease is high in India, this paper aims at exploring the probable image processing and data mining techniques that help in predicting the stage of cancer using images of cervical cells.

When a woman is suspected as the victim of cervical cancer, a Pap smear test can be performed for further diagnosis. Tissues are taken from the cervix and subjected to dyeing process to prepare a slide of cervical cells. Cervical cytology images cells are examined through a microscope by pathologists. These cells have three components as background, cytoplasm and nucleus. The morphological and texture features of nucleus and cytoplasm are adequate to decide the stage of cancer. Since manual diagnosis is error-prone and time consuming, an automated system employing image processing schemes is preferred. This paper attempts to use the computer aided image processing and data mining techniques on

* *Author for correspondence*

the RGB images obtained from the glass slide containing Pap smears and find out whether they are images of cancerous cells or normal cells.

Undesirable artifacts and background are removed using preprocessing techniques. The preferred components are extracted using segmentation methods. Size, shape and texture features of nucleus and cytoplasm are analyzed by data mining techniques and the malignant and benign cells are identified. This paper deals with the features that can be extracted from the cervical cells and selection of appropriate features for cell discrimination.

The rest of the paper is organized as follows: Section 2 deals with materials and methods of the work. Section 3 describes the results and discussion.

2. Materials and Methods

2.1 Cervical Cytology Images

Cervical cytology images from the database of Herlev hospital Denmark[3], are used. These are RGB color images in bitmap format. These images are grouped into seven categories by pathologists as given below:

- Normal Superficial
- Normal Intermediate
- Normal Columnar
- Light dysplastic
- Moderate dysplastic
- Severe dysplastic
- Carcinoma in situ

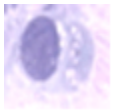
The first three categories are normal cells and the other four categories are abnormal or cancerous cells. Samples for each category are given in Table 1.

2.2 Preprocessing

2.2.1 Smoothening

The quality of the cervical cell image necessitates preprocessing to enhance features of regions of significance. RGB images of cervical cells taken from the microscope are preprocessed to remove blood and excessive dye stains. A Gaussian filter is convolved over the image to smoothen the regions of interest. Noise and other artifacts are suppressed. Edges of cytoplasm and nucleus are sharpened using a sharpening filter. This is achieved by subtracting a scaled unsharp version of the image from the original image.

Table 1. Samples of cervical cytology images of seven categories

Sl. No.	Various types of cell images: Types	Samples	Segmented Cytoplasm
1	Normal Superficial		
2	Normal Intermeidate		
3	Normal Columanr		
4	Light dysplastic		
5	Moderate dysplastic		
6	Severe dysplastic		
7	Carcinoma in situ		

2.2.2 Background Removal

The background information is not relevant for the diagnosis. So it can be suppressed. Fuzzy thresholding is used to find out the intensity of background[4]. The background is removed by using this threshold. The region of interest containing cytoplasm and nucleus is selected using active contours.

2.3 Segmentation

Two segmentation techniques *viz.*, Color K-means clustering and Gaussian Mixture Model are attempted on the image. In both cases, the number of clusters is determined using statistical methods. K-means clustering partitions the image into k clusters in which each pixel belongs to the cluster with the nearest mean intensity, serving as a prototype of the cluster. This results in partitioning of the image space into clusters of similar

intensity. The clusters having cytoplasm and nucleus information are detached from others and are given as inputs for feature extraction procedures.

Gaussian Mixture Model (GMM)[5,6] is another system utilized for bunching of nucleus and cytoplasm pixels. Pixels are fit into a GMM cluster by assigning them to the multivariate normal components that maximize their probabilities of belonging to that component. This method of assigning an input pixel to precisely one cluster is called *hard* clustering. Gaussian mixture model that employs Hopfield Markov Random field Expectation Maximization function is employed for segmentation of nucleus and cytoplasm of cells. GMM clustering can accommodate clusters that have varied sizes and correlation structures within them. Because of this GMM clustering supersedes *k*-means clustering. Comparison of K-means clustering and GMM methods were done based on segmentation metrics as Jaccard and Dice coefficients[7]. GMM surpasses K-means clustering in this comparison also. Clusters of cytoplasm and nucleus are stored as separate RGB images to ease feature extraction.

2.4 Feature Extraction

In image processing terminology, a feature refers to information chosen from the image, pertaining to the application. It may be a structural element or a quantifiable attribute of image. Morphological features like shape, size and textural features as first order, second order and third order moments of intensity, Gray Level Co-occurrence Matrix (GLCM) are extracted in this paper. Feature extraction involves reducing the amount of details required to describe a large set of data usually and especially an image in this context. Morphological and location features are calculated from the images of nucleus and cytoplasm.

2.5 Features

Zones of measurement in this work are nucleus and cytoplasm of cervical cell images. Morphology in biology refers to the study of form and structure of organisms and their specific structural features. Morphological features like size, shape and textural features are extorted as follows.

2.5.1 Size Features

a) Area of nucleus and cytoplasm : The number of pixels in the nucleus and cytoplasm respectively are counted to determine their area.

b) N/C ratio : The size of the nucleus with respect to cytoplasm determies the severity of cancer. The benign cells have small, round and compact nucleus located in the center of the cytoplasm. The cancer cells have a larger, irregularly shaped and sparse nucleus whose center deviates from the center of nucleus. The N/C ratio[8] is a foremost feature deciding the nature of cells.

$$N/C = Nucl_{area}/(Nucl_{area}+Cyto_{area}) \quad (1)$$

c) *Minor and Major axes length* : These measurements of the vital components of image decide the roundness of the nucleus and cytoplasm. The minor axis forms the smallest diameter and major axis forms the largest diameter.

2.5.2 Shape Features

a) *Perimeter of cytoplasm and nucleus:* The number of pixels that lie along the border are added up to from the perimeter.

2.5.3 Location Feature

The distance between the centroids of nucleus and cytoplasm is also used to determine the malignancy of cells.

2.5.4 Textural Features

Texture of an image is a set of metrics calculated from image to quantify the perceived spatial arrangement of color or gray intensities in an image or region of interest in an image.Textures are formed as an array of texture elements called as texels.Textural features enumerate the overall local intensity variability inside the object of interest.

a) Solidity : Solidity is a scalar that specifies the proportion of the pixels in the convex hull that are also in the region.

$$Solidity= Area/ConvexArea \quad (2)$$

Solidity of nucleus and cytoplasm are estimated.

b) Luminance : Luminance is the brightness perceived by human eye. It is calculated using the average intensity values of the three color channels.

$$Y=0.299 *Red_{\mu}+0.587 *Green_{\mu}+0.114 *Blue_{\mu} \quad (3)$$

2.5.5 Haralick Features

A co-occurrence matrix of an image is one that defines the co-occuring intensity values , either color or gray of an image, at a given offset. The size of this matrix is large,

increasing the time complexity of image processing. So Haralick[9] mined 14 metrics of this matrix along x and y directions, to form an useful feature set. The gray level co-occurrence matrix is used in this paper. All the metrics are calculated for GMM segmented cervical cytology images, but a few significant featues are mentioned below.

a) Angular Second Moment (ASM) :

$$tf_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\{p(i,j)\}^2 \tag{4}$$

b) Contrast:

$$tf_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g}\sum_{\substack{j=1 \\ |i-j|=n}}^{N_g} \{p(i,j)\} \right\} \tag{5}$$

c) Correlation :

$$tf_3 = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)*p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y} \tag{6}$$

d) Variance:

$$tf_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu)^2 p(i,j) \tag{7}$$

e) Inverse Difference Moment :

$$tf_5 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{1}{1+(i-j)^2}p(i,j) \tag{8}$$

f) Entropy:

$$tf_6 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}p(i,j)\log(p(i,j)) \tag{9}$$

g) Sum Entropy:

$$tf_7 = -\sum_{i=2}^{2N_g}p_{(x+y)}(i)\log\{p_{(x+y)}(i)\} \tag{10}$$

h) Difference Entropy :

$$tf_8 = -\sum_{i=0}^{N_g-1}p_{(x+y)}(i)\log\{p_{(x-y)}(i)\} \tag{11}$$

i) Information Measure of Correlation-1

$$tf_9 = \frac{HXY-HXY1}{\max\{HX,HY\}} \tag{12}$$

j) Information Measure of Correlation-1

$$tf_{10} = (1-e^{-2(HXY2-HXY)})^{1/2} \tag{13}$$

where

p(i,j) - (i,j)th entry in a normalised graytone spatial dependence matrix,

$p_x(i)$ – the ith entry in marginal-probability matrix obtained by summing the rows of p(i,j) ,

$$p_x(i) = \sum_{i=1}^{N_g}P(i,j) \tag{14}$$

$$HXY = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}P(i,j)\log(p(i,j)) \tag{15}$$

$$p(i,j)\log\{(p_x(i)p_y(j))]] \tag{16}$$

$$HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}P_x(i)p_y(j)\log\{p_x(i)p_y(j)\} \tag{17}$$

N_g is the number of distinct gray levels in the quantized image.

2.5.6 Tamura Features

Textural features[10] identified by Tamura correspond to human visual perception. Coarseness and directionality features are extracted from the image.

1. Coarseness : Coarseness refers to the roughness of a region. It tries to identify all possible textures in an image, irrespective of its size. It has a direct relationship to scale and repetition rates of varying grayscale distributions. Nucleus of malignant cells is coarser than normal cells.
2. *Directionality:* Two edge detectors are convolved over the image to detect edges in the image. At each pixel the angle and magnitude are calculated. A histogram of edge probabilities is built up by counting all points with magnitude greater than a threshold and quantizing by the edge angle. The histogram will reflect the degree of directionality. For images without strong edges the histogram is uniform. But for highly directional images, it shows peaks. The degree of directionality relates to the sharpness of the peaks.

2.6 Classification

Classification discriminates cervical cells into several groups based on homogeneous characteristics existing in them. These groups or classes are hence based on features of the cells. Classification may be done by supervised or unsupervised techniques.

Classification involves multiple steps:

The first step is defining groups. Number of groups in a classification is obviously application dependent.

Cervical cytology cells can be classified as normal or abnormal. Here the number of classes is two. Herlev database classifies the cells into seven categories as

1. Normal Superficial
2. Normal Intermediate
3. Normal Columnar
4. Light dysplastic
5. Moderate dysplastic
6. Severe dysplastic
7. Carcinoma in situ

There are seven classes of cells in this database.

The second step in classification is feature selection. Features that uniquely differentiate the cell images are mined to accelerate the classification process and to reduce the feature set used.

The third step is sampling of training data. The selection of training data decides the accuracy of discrimination and helps in design of decision rules.

The fourth step is estimation of universal statistics. Selection of decision rules is done by comparing training data with existing classification methods and choosing the closest matching scheme.

The final step is classification. It groups the images under the appropriate classes using the preferred classifier.

Classifiers like Support Vector Machine (SVM), Multilayer perceptron, Naïve Bayes which are proficient in discrimination of images into multiple classes, are used. Usually SVM is binary classifier, but it can also be extended as a multiclass classifier.

2.7 Feature Subset Selection

Feature selection[11–13] aims to derive a small subset of highly distinguishing features from a large set of extracted features. A good feature subset is one that contains features highly predictive of the class, but uncorrelated with each other. Feature selection optimizes time and space complexity for image classification. It helps in precise and efficient interpretation of information present in the images.

The feature selection process used in cervical cytology images must decide on a subset of features that enhance classification of cells as normal or cancerous for binary classifiers and classification into different stages of cancer in a multiclass classifier.

Correlation based feature subset selection (CFS) algorithm is used in this work. This algorithm relies on a heuristic for evaluating the merit of a subset of features. This heuristic considers the usefulness of individual features for predicting the class label along with the level of inter-correlation among the features. A feature is considered relevant if its value varies systematically with category membership. A feature is said to be redundant if one or more of the other features are extremely correlated with it. Subset of features with relevant features highly correlated with the output class while having low inter-feature correlation is selected by the CFS method. Dimensionality reduction is achieved by eliminating irrelevant features.

2.8 Classifiers

The attributes selected using the CFS algorithm[14] are used to classify the cells into binary or seven groups using the classifiers. Classifiers like multilayer perceptron, Support vector machine, BayesNet are used. A multilayer perceptron is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. Support Vector Machine (SVM) is a non probabilistic linear binary classifier and supervised learning model that assigns the input to any one of the two classes. Bayesian networks are used in medical diagnosis for representing the probabilistic relationships between symptoms and diseases. Binary sequential minimal optimization is another classifier using quadratic programming concept.

3. Experiment and Results

About 625 images from the Herlev hospital database are segmented by GMM and 100 features in total of cytoplasm and nucles are extorted. The database contains three groups of normal cells and four groups of cancerous cells. Data mining methods are used for classification of cells in two ways as malignant and benign cells and the seven types as mentioned in c.

3.1 Feature Selection

Classification is attempted using all features with Multilayer perceptron model. The time taken for building the model is 18 seconds and for classification is 15 minutes. So subset of features is derived for efficiency.

Out of the 100 features extracted, 10 features are selected using Correlation based feature subset selection method for the binary classification.

1. N/C ratio
2. Length of Minor axis of nucleus
3. Length of Major axis of cytoplasm
4. Length of Minor axis of cytoplasm
5. Solidity of cytoplasm
6. Contrast of cytoplasm
7. Contrast of nucleus
8. Perimeter of nucleus
9. Cluster prominence of cytoplasm
10. Autocorrelation of nucleus

The time taken for building the model in this case is 2 seconds and for classification is 20 seconds.

For the seven group classification, 11 features are selected using the same CFS method. They are

1. N/C ratio
2. Length of Minor axis of nucleus
3. Length of Major axis of nucleus
4. Length of Minor axis of cytoplasm
5. Length of Major axis of cytoplasm
6. Contrast of cytoplasm
7. Solidity of cytoplasm
8. Distance between centroids of nucleus and cytoplasm
9. Perimeter of cytoplasm
10. Autocorrelation of nucleus
11. Luminance of nucleus

3.2 Results

Tables 2 shows the results of binary grouping i.e., classification of cervical cells into normal and cancerous cells. Table 3 depicts the results for classification into seven types using different classifiers.

{Place Table 2 here}

Table 2. Kappa statistics for binary classifiers

Classifier	Correctly classified instances	Incorrectly classified instances	Kappa statistic
Multilayer Perceptron	625	0	1
Binary SMO	624	1	0.9968
NaiveBayes	610	15	0.9515
BayesNet	624	1	0.9968

Table 3. Classification as seven groups

Classifier	Correctly classified instances	Incorrectly classified instances	Kappa statistic
Multilayer Perceptron	551	74	0.8535
Attribute selected classifier	547	78	0.8474
Bagging	489	136	0.7286
LibSVM	625	0	1

Kappa statistics is a chance-corrected measure of agreement between the classifications and the true classes. Kappa statistics is calculated by the following equation.

$$k = \frac{Observed\ agreement - Agreement\ expected\ by\ chance}{Maximum\ possible\ agreement} \quad (18)$$

It is evaluated for all the classifiers. If Kappa statistics is greater than 0, it implies that the selected classifier is performing better than chance.

In medical image classification, statistical measures like true positive rate[15], false positive rate and precision play a vital role in judging the accuracy of the outputs. These measures are evaluated by the following equations.

$$True\ Positive\ Rate(TPR) = \frac{TP}{TP + FN} \quad (19)$$

$$True\ Negative\ Rate(TNR) = \frac{TN}{TN + FB} \quad (20)$$

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

where

TP – True Positive - Cancer cells identified as cancer cells

FP – False Positive – Normal cells incorrectly identified as cancer cells

TN – True Negative – Normal cells identified as normal cells

FN – False Negative – Cancer cells identified as normal cells.

Values calculated from the above equations for the classifiers are listed in Table 4 and Table 5. When the true positive rate is approximately 1, the classification has correctly identified cancer cells. When the true negative rate is about 0, the classification has correctly identified normal cells. The classifiers classify the inputs using the features selected by the CFS algorithm. They calculate the probable values of features to assign them to a class.

Table 4. Statistical measures for two group classifiers

Classifier	TPR	TNR	PRECISION
Multilayer Perceptron	1	0	1
Binary SMO	0.996	0	1
NaiveBayes	0.975	0.023	0.972
BayesNet	0.996	0	1

Table 5. CStatistical measures for seven groups classifiers

Classifier	TPR	TNR	PRECISION
Multilayer Perceptron	0.829	0.022	0.903
Attribute selected classifier	0.967	0.042	0.85
Bagging	0.797	0.046	0.81
LibSVM	1	0	1

4 Discussion

Feature set selection and classification of cells done by data mining techniques like feature subset selection and classification is found to be efficient. There is a considerable reduction in the time and space required for classification. Due to the dimensionality reduction of the feature set. Values of true positive rate, false positive rate and precision show the efficiency of classifiers in predicting the group members. Multilayer perceptron works well for the binary classification and Library Support Vector Machine suits the seven-way classification well.

5. Conclusion

In this paper, an enhanced method for classifying Pap smear images using selected feature set is proposed. The Pap smear images are pre-processed, segmented and classified for diagnosis of cancer. Size, shape and textural features of cervical cells decide the stage of cancer. These features are extracted from the nucleus and cytoplasm of the cells. The classifiers like Multilayer perceptron, Support vector machine, BayesNet are used to classify the cells based on the subset of features extracted using the data mining technique called Correlation based feature subset selection algorithm. The outputs of the classifiers are analyzed based on Kappa statistics, true positive rate, true negative rate and precision. Multilayer perceptron and Support Vector Machine show promising results. This study reveals that binary or multiclass grouping can be done effectively by data mining techniques. The sets of features selected are class dependent. Multilayer perceptron and SVM are useful in majority of classifications.

6. References

1. National Cancer Institute – United States, Seer Stat Fact Sheets: Cervix Uteri Cancer. Available from: http://seer.cancer.gov/statfacts/ html/cervix.html. Date accessed: 01/02/2016
2. ICO Information Centre on HPV and Cancer (HPV Information Centre), Human Papillomavirus and Related Diseases Report. Available from: http://www.hpvcentre.net/IND.pdf Date accessed: 21/12/2015
3. Pap-smear (DTU/Herlev) databases & related studies. Hervelv Univerity Hospital Database. Available from: http://labs.fme.aegean.gr /decision/downloads Date accessed: 20/10/2011
4. Lakshmi GK, Krishnaveni K. Automated extraction of cytoplasm and nuclei from cervical cytology images by fuzzy thresholding and active contours. International Journal of Computer Applications. 2013 Apr; 73(15):26–30.
5. Lakshmi GK, Krishnaveni K. Multiple feature extraction from cervical cytology images using Gaussian mixture model. Proceedings of World Congress on Computing and Communication Technologies; India. IEEE; 2014. p. 309–311. Doi: 10.1109/WCCCT.2014
6. HuangY. Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. IEEE Transcations on Biomedical Engineering. 2005 Nov; 52(11):1801–11.
7. Lakshmi GK, Krishnaveni K. Quantitative validation of segmentation methods of cervical cytology images. International Journal of Innovations in Engineering and Technology (IJIET). 2015 Dec; 6(2):75–80.
8. Mahanta LB, Bora K. Analysis of malignant cervical cells based on N/C ratio using pap smear images. International Journal of Advanced Research in Computer Science and Software Engineering. 2012 Nov; 2(11):341–6.
9. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transaction on Systems, Man and Cybernetics. 1973 Nov; 3(6):610–21.
10. Rodenacker K, Bengtsson E. A feature set for cytometry on digitized microsopy images. Analytical Cellular Pathology. 2003. p. 1–36.
11. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. 1–33. Available from: http:// www.public.asu.edu/~jtang20/publication/feature_selection_for_classification.pdf. Date accessed: 03/04/2014.
12. Nixon, Aguado AS. Feature extraction and image processing for computer vision. Oxford. Newnes Press: 2002. p. 1–336.
13. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics. 2015; 15:13.
14. Veeraiah D, Vasumathi D. Feature sub selection over high dimensional data based on classification models. Indian Journal of Science and Technology. 2016 Feb; 9(8):1–6.
15. Andreas GK, Wilfried NJ. Gansterer on the relationship between feature selection and classification accuracy. Journal of Machine Learning. 2008; 4(1):90–105.