Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/98380, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Classification of Clinical Dataset of Cervical Cancer using KNN

Meenakshi Sharma*, Sanjay Kumar Singh, Prateek Agrawal and Vishu Madaan

School of Computer Science Engineering, Lovely Professional University, Phagwara - 144411, Punjab, India;
meenakshis84@gmail.com, sanjayksingh.012@gmail.com, prateek061186@gmail.com,
vishumadaan123@gmail.com

Abstract

Background/Objectives: The primary objective of this paper is to classify the clinical dataset of cervical cancer to identify the stage of cancer which helps in proper treatment of patient suffering from cancer. **Methods/Statistical Analysis:** This research work basically moves toward the detection of cervical cancer using Pap smear images. Analysis of Pap smear of cervical region is an efficient technique to study any abnormality in cervical cells. The proposed system firstly segment the pap image using Edge Detection to separate the cell nuclei from cytoplasm and background and then extract various features of cervical pap images like area, perimeter, elongation and then these features are normalized using min-max method. After normalization KNN method is used to classify cancer according to its abnormality. **Findings:** The classification accuracy with 84.3% of maximum performance with no validation and classification accuracy with 82.9% of maximum performance with 5 Fold cross validation is achieved.

Keywords: Cervical Cancer, Cell Images, Classification, Pap Smear Test

1. Introduction

In developing countries cancer of cervix afflicts most of the women[1]. Root cause of cervical cancer is HPV (Human Papilloma Virus) infection. Effect of cancer is avertable if detected in early stage because growth of cervical cancer took 10-15 years. For early detection of Cervical Cancer preliminary test was introduced by[2] named as Pap Test. This is an effective test to identify pre-cancerous lesions of cervical cancer.

In past few years, based on deep study of pap images various methods were developed by various researchers to automate detection process. In past benchmark is provided for the classification of cervical cancer[3]. Various approaches have been developed for this purpose, such as classification model based on OLMAM and LMAN[4], H_2MLP model[5,6], Debris removal from Pap images[7], colored based watersehed[8], Fuzzy based classification[9,10] and various approches for segmentation[11-16].

2. Proposed Methodology

2.1 Data Collection and Pre-Processing

Dataset used in this work is collected from Fortis Hospital Mohali, Punjab (India).This data set include images acquired through digital camera attached with microscope. Image preprocessing basically deals with enhancing image quality for proper vision. In proposed method Gaussian filter is used to remove unwanted noise and Histogram equalization is used to enhance image quality as shown in Figure 1a, Figure 1b and Figure 1c.

2.2 Segmentation of Pap Images

Image segmentation is basically used to extract the cell nucleus which is region of interest for our study. Image segmentation basically separate cell nuclei from cytoplasm which is further used for features extraction. In proposed method edge based segmentation is used to extract desired cell nuclei. Segmentation method that is

*Author for correspondence

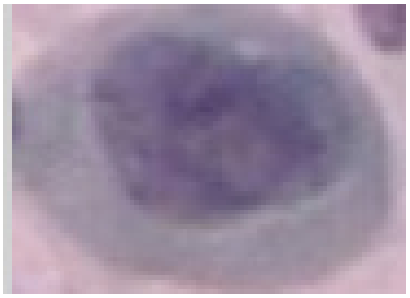

Figure 1(a). Original cervical pap image.

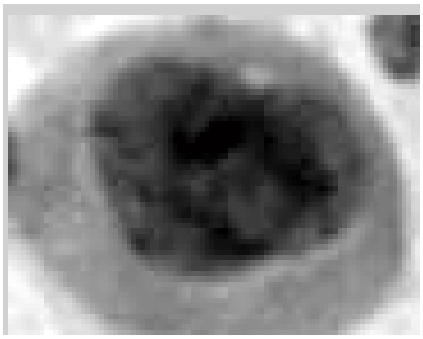

Figure 1(b). Enhanced gray scale image.

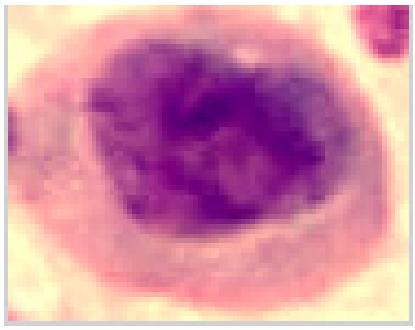

Figure 1(c). Enhanced RGB image.

based on edge detection follows a method to determine the relationship exists between pixels of neighborhood[17]. Convolution of Gradient operator with image is used to perform edge detection. Edge detection basically marks the boundaries of particular image. In edge detection process when the magnitude of gradient operator exceeds the threshold value then particular edge is detected.

In our research work for detecting edge of pap image sobel gradient operator is used. Sobel operator basically contains two 3*3 kernels where the original image convolves with these two kernels to calculate derivatives for vertical and horizontal. Process of sobel operator work as follows where Ga and Gb are two derivatives and I is the original image and G is the gradient magnitude as shown in Equation 1 and Equation 2[17].

$$G_a = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -3 & 0 & 3 \end{bmatrix} * I \quad G_b = \begin{bmatrix} -1 & -2 & -3 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \end{bmatrix} * I \qquad (1)$$

$$G = \sqrt{Ga^2 + Gb^2} \qquad (2)$$

For image segmentation firstly we identify edges of pap image then we outline desired cell nuclei and then we finally extract desired cell nuclei as shown in Figure 2(a), Figure 2(b), Figure 2(c), Figure 2(d).

2.3 Features Extraction and Selection

For classification purpose, it is required to extract an optimal amount of features. By using huge amount of features computational overhead is increased. So it becomes necessary to extract optimal number of features used for classification. In our work various morphological features are extracted like Area, Elongation, Perimeter of cell nuclei and cytoplasm and finally N/C ratio as shown in Table 1.

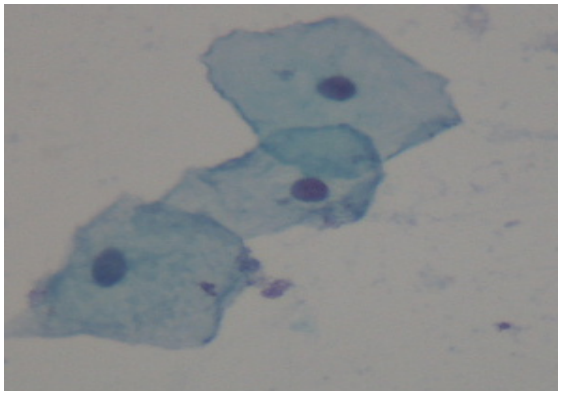

Figure 2(a). Original cervical pap image.

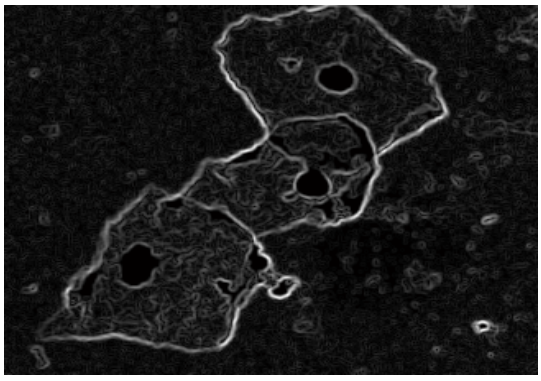

Figure 2(b). Edge detection of pap image.

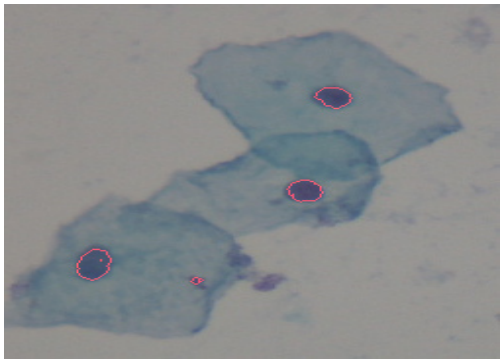

Figure 2(c). Outline cell nuclei.

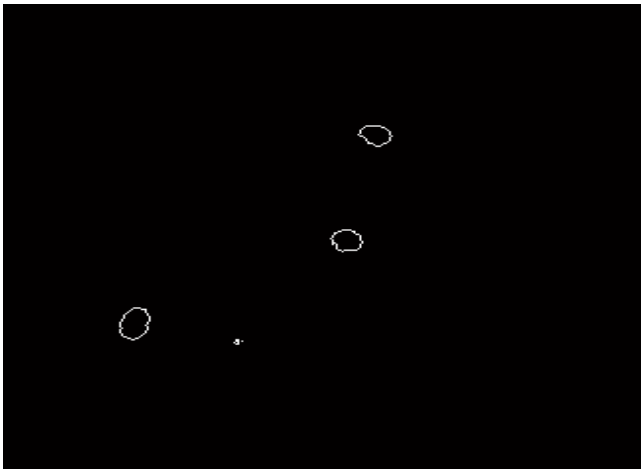

Figure 2(d). Extracted cell nuclei.

Table 1. Extracted features

Extracted Features						
Nucleus Area	Cytoplasm Area	Nucleus Perimeter	Cytoplasm Perimeter	Nucleus Elongation	Cytoplasm Elongation	N/C Ratio
3816.5	7560.5	225.25	467.375	0.999271	0.8590816	0.335458
2370	5271.875	176	429.75	1.057293	0.45882224	0.310133
2268.5	3691.5	175.125	273	0.825098	1.17770435	0.380621
3816.5	7560.5	225.25	467.375	0.999271	0.8590816	0.335458
3816.5	7560.5	225.25	467.375	0.999271	0.8590816	0.335458
2268.5	3691.5	175.125	273	0.825098	1.17770435	0.380621
4221	3650.75	238.375	342.375	0.97529	0.98688525	0.536221
2564.375	5768.25	194.125	399.875	0.809412	0.67346752	0.307751

After calculating various features of cervical cells we normalize the extracted features using min-max approach.

2.4 Classification of Pap Images

K-Nearest Neighbor is a non-parametric classifier that is used for regression and classification. The word parametric means that we cannot make assumptions on the data distribution. In KNN classifier, there is no need of explicit training phase. Centroid is just a center of particular cluster as shown in Figure 3.

In KNN the data is divided into test set and training set. For every row of test set nearest neighbor k based on Euclidean distance of training set point are observed and based on the majority votes classification is achieved. For selecting K nearest point from the data of training set, Euclidian distance is measured based on following Equation 3, 4 and 5[17,18]:

$$a_i = \{a_i1, a_i2, a_i3, a_i4 \ldots\ldots\ldots\ldots\ldots\ldots\ldots a_in \quad (3)$$

$$b_i = \{b_i1, b_i2, b_i3, b_i4 \ldots\ldots\ldots\ldots\ldots\ldots\ldots b_in \quad (4)$$

$$ed_i = \|b_i - a_i\| = \left|\sum_{j=1}^{n} \frac{|b_{ij} - a_{ij}|^2}{n}\right| \quad (5)$$

Where a_i and b_i are training and testing data set and n is total number of features for each sample.

Table 2 describes basic outcome KNN classifier with no cross validation status. Figure 4 and Figure 5 shows scatter chart with outcome of KNN classifier and it shows confusion matrix of KNN with no cross validation.

Table 3 describes basic outcome KNN classifier with 5 fold cross validation status.

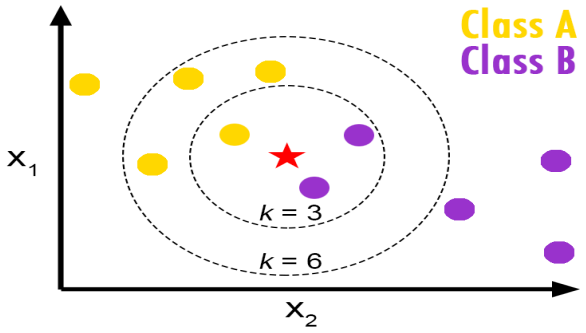

Figure 3. Basic approach of KNN Classifier[17,18].

Table 2. Output of KNN with no validation

Classifier	Accuracy	Prediction Speed	Training Time	Distance Metric	No. of Features	PCA Status	Validation
Fine KNN	84.3%	~3700 obs/sec	0.29412 secs	Euclidian	All Features	Enable	No Cross Validation

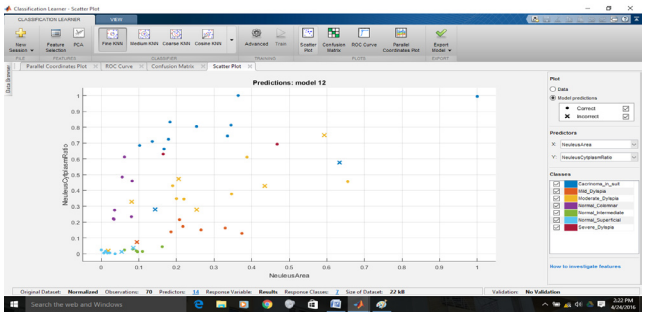

Figure 4. KNN with no validation.

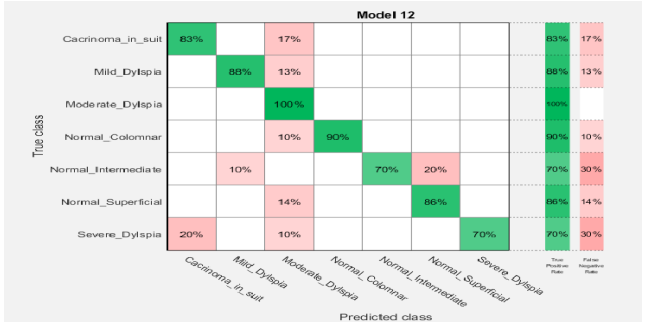

Figure 5. Confusion matrix for KNN with no validation.

Table 3. Output of KNN with validation

Classifier	Accuracy	Prediction Speed	Training Time	Distance Metric	No. of Features	PCA Status	Validation
Fine KNN	82.9%	~840 obs/ sec	0.57709 secs	Euclidian	VarName1, VarName3	Enable	5 Cross Validation

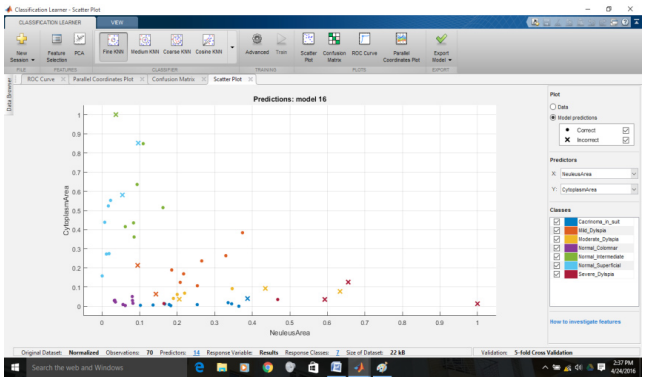

Figure 6. KNN with validations.

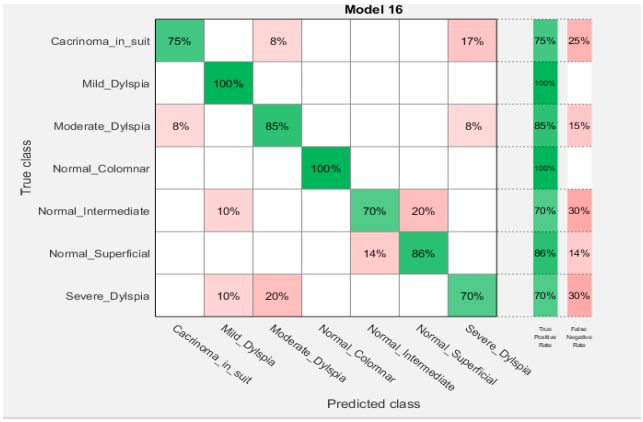

Figure 7. Confusion matrix for KNN with validation.

Figure (6), Figure (7) shows scatter chart with outcome of KNN classifier and it shows confusion matrix of KNN with cross validation.

3. Conclusion and Future Scope

In this paper segmentation approach efficiently separate the cell nuclei from its cytoplasm. A variety of features are extracted from the pap images which describe the characteristics of cell nuclei and cytoplasm. We concluded that KNN has given classification accuracy with 84.3% of maximum performance with no validation and classification accuracy with 82.9% of maximum performance with 5 Fold cross validation.

Furthermore, our work showed the classification of cervical cancer according to their anomaly degree by using only limited features with more number of features we can increase the accuracy of classification.

4. Acknowledgment

I would like to convey my gratefulness to Dr. Ritu Pankaj, Associate Consultant SRL Pathology department of Fortis Hospital Mohali (Punjab), India for providing the pap images. This paper has the assent of all co-authors and authorities of my institute, where this study has been carried out and there exists no conflict of interest anywhere.

5. References

1. American Cancer Society. Cancer Facts and Figures. 2015. Available from: http://www.cancer.org/acs/groups/cid/documents/webcontent/003094-pdf.pdf

2. Papanicolaou GN. The cell smears method of diagnosing cancer. American Journal of Public Health. 1948 Feb; 38(2):202–5.
3. Byriel J. Neuro-fuzzy classification of cells in cervical smears. Denmark, Oersted. 1999; 8(1):38.
4. Plissiti ME, Charchanti A. Automated segmentation of cell nuclei in PAP smear images. ITAB Proceedings of International Special Topic Conference on Information Technology in Bio Medicine; Greece, Ioannina. 2006.
5. Ampazis N, Dounias G, Jantzen J. Pap-smear classification using efficient second order neural network training algorithms. Lecture Notes in Artificial Intelligence; 2004 May. p. 230–45.
6. Othman NH. Capability of new features of cervical cells for cervical cancer diagnostic system using hierarchical neural network. IJSSST. 2008; 9(2).
7. Malm P, Balakrishnan N, Sujathan VK, Kumar R, Bengtsson E. Debris removal in Pap-smear images. Computer Methods and Programs in Biomedicine. 2013 Jul; 111(1):128–38.
8. Lezoray O, Cardot H. Cooperation of color pixel classification schemes and color watershed: A study for microscopic images. IEEE Trans Image Process. 2002 Jul; 11(7):783-9.
9. Begelman G, Gur E. Cell nuclei segmentation using fuzzy logic engine. ICIP Proceedings of International Conference on Image Processing; 2004 Nov. p. 2937–40.
10. Hiremath PS. Fuzzy Rule based classification of microscopic images of squamous cell carcinoma of esophagus. International Journal of Computer Application. 2011 Jul; 25(11):30- 3.
11. Lee KM, Street WN. Learning shapes for automatic image segmentation. Proceedings of INFORMS-KORMS Conference; Seoul Korea. 2000 Jun. p. 1461-8.
12. Bamford P, Lovell B. A water immersion algorithm for cytological image segmentation. Proceedings of the APRS Image Segmentation Workshop; 1996. p. 75-9.
13. Costa JAF, Mascarenhas NDA. Society for optical engineering cell nuclei segmentation in noisy images using morphological watersheds. International Proceedings on PIE of 3164; 1997. p. 314-24.
14. Mouroutis T, Roberts S, Robust J. Cell nuclei segmentation using statistical modeling. IOP Bioimaging. 1998 Jun; 6(2):79-91.
15. Bamford P, Lovell B. Unsupervised cell nucleus segmentation with active contours. Signal Processing. 1998 Dec; 71(2):203-13.
16. Garrido A, Blanca PDLN. Applying deformable templates for cell image segmentation. Pattern Recognition. 2000; 33(5):821-32.
17. Yi-Wei Y, Jung-Hua W. Image segmentation based on region growing and edge detection. IEEE SMC Conference on Systems, Man and Cybernetics; 1999 Oct. p. 798-803.
18. Gonzalez W. Digital Image Processing using Matlab. Gatesmark Publishing; 2009. p. 827.