Cervical Cancer Cell Identification & Detection Using Fuzzy C Mean and K nearest Neighbor Techniques

Bhuvaneshwari K V, Poornima B

Abstract - Across the globe, woman has been diagnosed two major forms of cancer, in which one is identified as cervical cancer and its micro classification. Morphology changes in cells or dead nucleus in the cervix causes cervical cancer. These cells are characterized with multiple nucleuses, faulty & lack of cytoplasm and so on. Detection of cervical cancer using smear test is extremely challenging because such cells does not offer texture variations or any significant color from the normal cells. Therefore for identification in abnormality of cells we required high level Digital image processing technique which compromises an automated, comprehensive machine learning skills. An advanced Fuzzy based technique has been implied to separate nucleus and cytoplasm from the cell. KNN is instructed with the color features and shape features of the segmented units of the cell and then an unknown cervix cell samples are classified by this technique. The proposed technique gives shape and color features of nucleus and cytoplasm of the cervix cell.

Key Words: KNN, Fuzzy C mean (FCM), Gustafson-Kessel (GK) Clustering, Gray Level Co-Event Matrix(GLCM)

I. INTRODUCTION

Early and adequate treatment is most important in prevention of pre-cancerous cells to develop. By using Georges Papanicolaou pap-smear technique, cyto-technicians are possible to detect pre-cancerous cells in uterine cervix by staining the cells using microscope. Figure 1 depicts affected cervix cell. Department of Pathology helps in classification of specimen by specimen with the help of well-trained cyto-technicians using microscope. This method does not provide actual solution since possibilities of human error rate. To overcome from this need a computational technique such as image processing and machine learning.

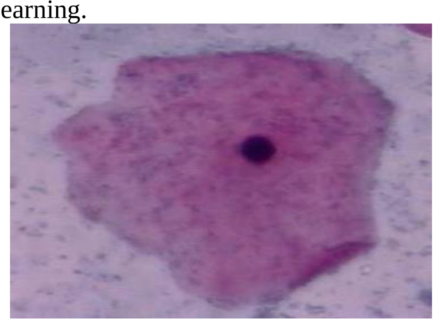


Fig 1: Singe pap-smear cell image.

Image preparing is a technique to change over a picture into computerized structure and play out certain tasks on it, so as to get an upgraded picture or to extricate some helpful data from it. It is used to extract the different features like area, color, shape which helps to distinguish malignant and benign tumor. To identify the region of interest in a tumor, first acquire the images like microscopic, pap smear, CT scan and MRI and by applying suitable pre-processing and segmentation method features are extracted.

Artificial intelligence is a part of man-made brainpower that utilizes an assortment of factual, probabilistic and streamlining procedures that enables PCs to gain from past precedents and to identify new examples from huge, boisterous or complex dataset. It is an integral asset for therapeutic picture examination and is every now and again utilized in cancer diagnosis and recognition by means of pictures.

It is used to develop a model by observing the features of training dataset. This model is used to predict the unseen data. The model can be developed using many supervised and unsupervised learning algorithms. The AI procedure spares the time and exertion expected to find the example or to build up a grouping plan.

1.1 Background

Herlev University Hospital has developed two data sets related to cervical cell by the cyto-technicians for classification purposes. Martin uses the feature extraction of image processing in Matlab. He compares with classification results of new data with old data, and concludes a desirable error rate for both false-negative (FN) and for the false-positive (FP) as well. The results depicts false-positive error is significantly larger than the false-negative. In medical terms a bad smear screening is referred to as a positive test result. So positive cells mean abnormal cells and negative cells mean normal cells. The main methods used by Martin were Hard C-means (HCM), Fuzzy C-means (FCM) and Gustafson-Kessel clustering (GK). Finding the optimal fuzzy exponent for FCM was one of the investigations. Also direct and hierarchical classifiers were examined, but no linear classifier was tested.

II. METHODOLOGY

Fig 2 illustrates the system architecture of the proposed work:

Manuscript published on 30 August 2019.
***Correspondence Author(s)**

Bhuvaneshwari K V, Research Scholar, ISE Department, Bapuji Institute of Engineering & Technology, Davangere, Karnataka.
Poornima B, Professor and Head of ISE Department, Bapuji Institute of Engineering & Technology, Davangere, Karnataka.

Retrieval Number: I7892078919/19©BEIESP
DOI: 10.35940/ijitee.I7892.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

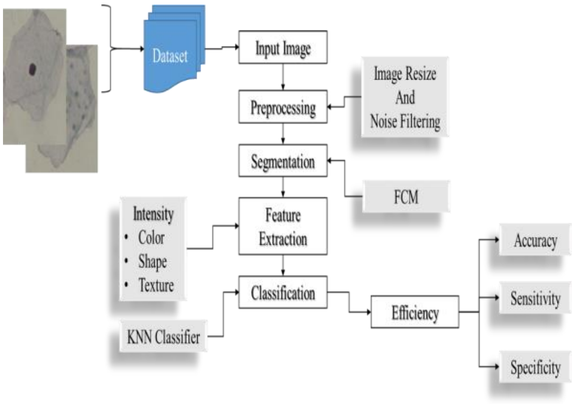


Fig 2:System Architecture

Input Image:

The single cell microscopic image data is collected from the cancer registry hospitals. Each image is trained individually. Then image dataset is divided into training and testing dataset. A image is a rectangular exhibit of pixels. Every pixel speaks to the estimation of some property of a scene estimated over a limited area. The property could be numerous things, yet generally measure either the normal dimensions (one esteem) or the brilliance of the picture separated through red, green and blue channels (three qualities). The qualities are regularly spoken to by an eight piece whole number, giving a scope of 256 levels of brightness. Information pictures are gathered gotten from the organizer. We can fetch all at a time but to obtain feature of individual image we will fetch one by one.

Pre-processing

Input image is fetched and resized. It is necessary to make the consistency in features for trained images and testing images of different size. So image will be resized to standard size as per user desire. In this we have resized the image to 256*256.

For Image resizing, Gaussian filtering is applied. Image blurring is done using Gaussian blur or smoothing with Gaussian function. To reduce the intensity of the pixel and image noise this function is used in graphics software. To enhance image structures, Gaussian smoothing is used in computer vision algorithms as a pre-processing stage at different scale of image structures enhancing. In this proposed work, Gaussian low pass filter is applied with 0.5 intensity to enhance input image.

Segmentation

In the proposed work Fuzzy C-implies bunching calculation is utilized for division. At present, the FCM calculation has been widely utilized in highlight investigation, design acknowledgment, image processing, classifier structure, and so forth. In any case, the FCM bunching calculation is delicate to the circumstance of the introduction and simple to fall into the nearby least or a seat moment that emphasizing. In numerous reasonable applications the bunching technique utilized is FCM with different restarts to escape from the reasonableness of the values.

The algorithm is repetition of cluster method that produces an optimal c partition by reducing the weighted sum of squared error objective function. A solution of the object function can be obtained via an iterative process, which is carried out as follows:

1. Set values for c, q and ϵ. Where c is number of clusters, q is a weighting exponent and ϵ is a measure of accuracy.
2. Fuzzy partition matrix U = [uik] should be initialized
3. Initialize loop counter b=0.
4. Calculate c cluster centres {vi(b)} with U(b) :

$$\mathrm{Vi(b)}=\frac{\sum_{k=1}^{n}(u_{ik}^{(b)})^{q}\,x_k}{\sum_{k=1}^{n}(u_{ik}^{(b)})^{q}}$$

5. Membership u(b+1) should be Updated

$$u_{ik}^{(b+1)}=\frac{\left(\left[\frac{1}{x_k-\mathrm{Vi}}\right]\right)^{1/(q-1)}}{\sum_{j=1}^{c}\left(\left[\frac{1}{x_k-\mathrm{Vi}}\right]\right)^{1/(q-1)}}$$

6. If ‖U (b) − U (b+1)‖ < ϵ, stop; otherwise, set b = b + 1 and go to step 4

Feature Extraction

Gray Level Co-event Matrix (GLCM) is utilized to figure the extraordinary reliance of dim dimensions in a picture. In GLCM the quantity of lines and segments are actually equivalent to the quantity of gray dimensions in the picture. Co-occurance networks are built in four spatial directions (0o,, 45o,90o,135o). Another network is built as the normal of going before frameworks. Let the Co-occurance lattice be Pij and the span of the framework is NxN.Each element (i,j) represents the frequency by which pixel with grey level i is spatially related to pixel with grey level j.

Table 1. Formulas to calculate Texture Features from GLCM

Sl.No	GLCM Feature	Formula
1.	*Contrast*	$\sum_{i,j=0}^{N-1} P_{i,j}\,(i-j)^2$
2.	*Correlation*	$\sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right]$
3.	*Dissimilarity*	$\sum_{i,j=0}^{N-1} P_{i,j}\,\lvert i-j\rvert$
4.	*Energy*	$\sum_{i,j=0}^{N-1} P_{i,j}^2$
5.	*Entropy*	$\sum_{i,j=0}^{N-1} P_{i,j}\,(-\ln P_{i,j})$
6.	*Homogeneity*	$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}$
7.	*Mean*	$\mu_i=\sum_{i,j=0}^{N-1} i\,(P_{i,j})\ ,\ \ \mu_j=\sum_{i,j=0}^{N-1} j\,(P_{i,j})$
8.	*Variance*	$\sigma_i^2=\sum_{i,j=0}^{N-1} P_{i,j}\,(i-\mu_i)^2\ ,\ \sigma_j^2=\sum_{i,j=0}^{N-1} P_{i,j}\,(j-\mu_j)^2$
9.	*Standard Deviation*	$\sigma_i=\sqrt{\sigma_i^2}\ ,\ \ \sigma_j=\sqrt{\sigma_j^2}$

Shape Features

Shape, an extent of the object is represented in the form of Binary. Calculation of external boundary and object's Geometric properties refers to Shape features.

Major shape feature calculated in our method are Circularity, Perimeter and Area. For an image classification, the meaningful information can be extracted from shape features. For various tumor types, the feature of the shape varies. Using image’s connected regions calculation of Shape features are done. For calculations of Area and Perimeter Boundary pixels are used. Perimeter and Area are utilized to determine Circularity. Limit for the pixels of the region is Ed (i,j).

Table 2- Formulas to calculate Shape Features using connected regions.

Sl.No	Shape Feature	Formula
1.	*Perimeter (P)*	$\sum_{i,j=0}^{M,N} E_d(i,j)$
2.	*Area (A)*	$\sum_{i,j=0}^{M,N} b(i,j)$
3.	*Circularity (C)*	$\frac{4\pi A}{p^2}$

Classification

k-Nearest Neighbors classifier is one of the easiest order procedures. K nearest preparing vectors are resolved for an information include vector X in the process as indicated by an appropriate separation metric. The vector X is alloted the class to which most of this k closest neighbors has a place with. The k-NN calculation depends on a separation and a casting a ballot work in k closest capacities. Euclidean separation is the measurement utilized in the classifier. k-NN calculation comprises of two stages one is preparing stage and other one testing stage. In preparing stage, the information focuses are encouraged in a n-dimensional space where n can be any positive number more noteworthy than 1. These preparation information focuses have names related with them that name their class. In testing stage, unlabelled information called information focuses are given and the calculation creates the closest information focuses to the unlabelled point. The Nearest Neighbor classifier chooses a solitary neighbor in the informational index and continues further while the k-Nearest Neighbors classifier chooses k neighbors and afterward takes most of the informational collection type as the outcome for the experiment. In this way, it gives great execution to ideal estimations of k. At long last, the calculation restores the class of larger part of that rundown. k-nearest neighbors algorithm is as follows:

1) Seat a proper distance metric.

2) In the training phase: Stores all the training data set P in pairs (according to the selected features) P = (yi, ci), i=1. . .n, where yi is a training pattern in the training data set, ci is its corresponding class and n is the amount of training patterns.

3) During the test phase: Computes the distances between the new feature vector and all the stored features (training data).

4) The k-nearest neighbours are chosen and asked to vote for the class of the new example. The correct classification given in the test phase is used to assess the correctness of the algorithm. If this is not satisfactory, the k value can be tuned until a reasonable level of correctness is achieved.

III. RESULTS AND DISCUSSION

Proposed technique has been implemented in the MATLAB 2017. Results are mentioned below.

Step 1: Each individual sample is selected for process as shown in Fig 3.

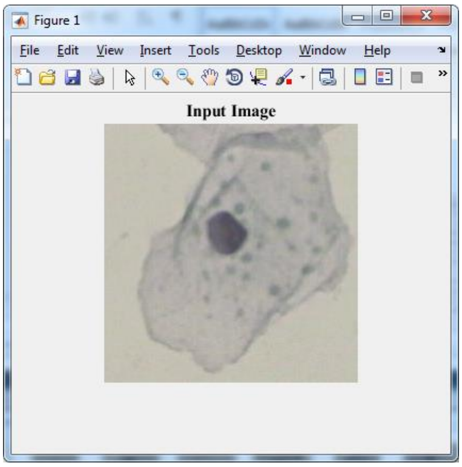


Fig 3: Input Image Selected

Step2: Input image is resized to 256*256, Gaussian filter is applied and then fuzzy c means segmentation is done as shown in Fig 4.

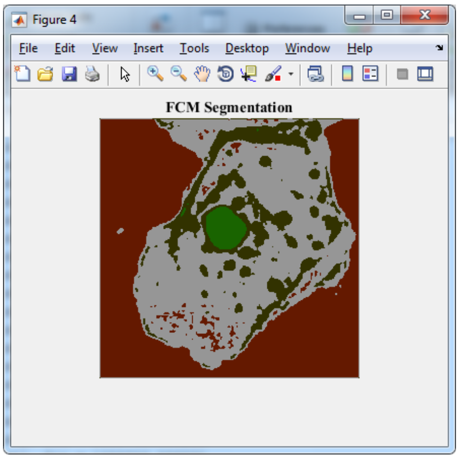


Fig 4: Fuzzy C Means Segmentation

Step3: For the fuzzy clustering method we obtain 3 clusters. As shown in Fig 5, index 1 refers to color changed segmented part, Index 2 refers to foreground image, Index 3 refers to background image and Index 4 is the original binary image. We have to select the segmented index. The appearance will change for each execution i.e., different images may appear on different index.

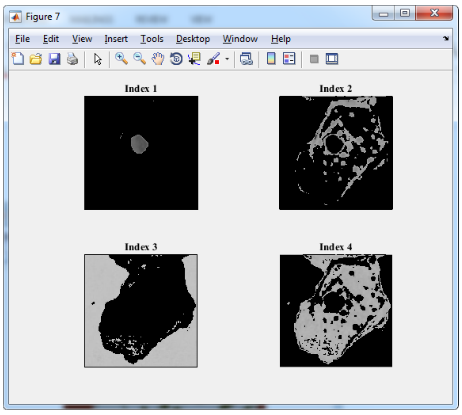


Fig 5.: Fuzzy clusters

Step 4: For each individual sample image collected text, shape and histogram features as shown in Fig 5.5. It also shows the feature values for selected image. Feature values vary from image to image. In this work we use GLCM feature extraction technique which gives 13 features. GLCM feature extraction is comparably best feature extraction technique for medical images. Along with GLCM features we also consider Histogram and Shape features. The reason for taking more features is to avoid the false detection

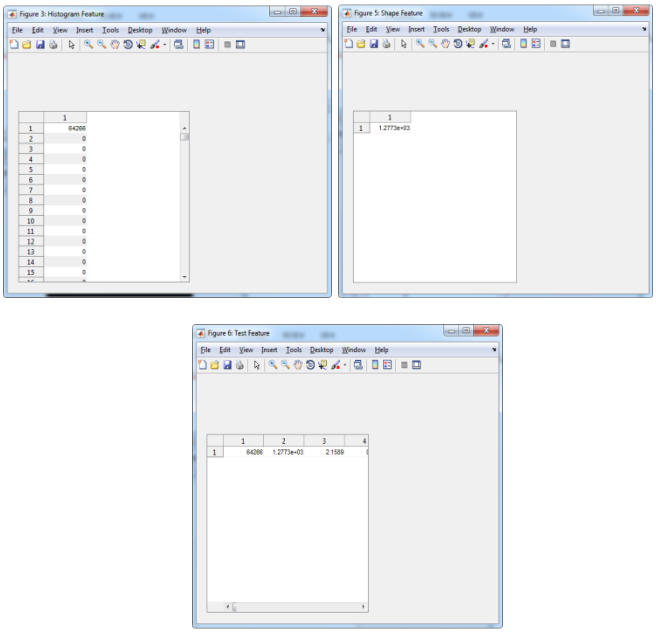


Figure 3: Histogram Feature
Figure 5: Shape Feature
Figure 6: Test Feature
File Edit View Insert Tools Desktop Window Help

	1
1	64266
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0

	1
1	1.2773e+03

	1	2	3	4
1	64266	1.2773e+03	2.1589	[illegible]

Fig 6: Text, Shape and Histogram features

Step 5: The 15 sample images are trained and 5 images are taken as test samples. The k-NN classifier is used as classification algorithm. It classifies the sample as shown in Fig 7.

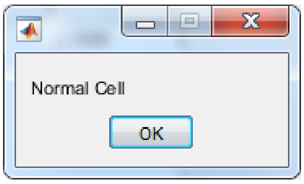
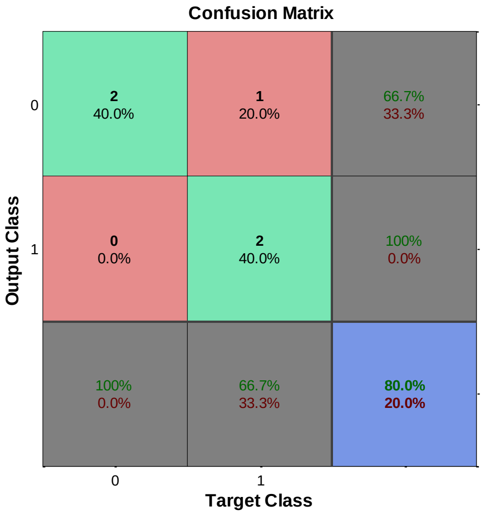

Normal Cell

Fig 7: Classification using k-NN

Step 6: To measure the performance of classifier we used different measurement coefficients such as accuracy, sensitivity, specificity using confusion matrix as shown in Fig 8.



	0	1	
0	2 40.0%	1 20.0%	66.7% 33.3%
1	0 0.0%	2 40.0%	100% 0.0%
	100% 0.0%	66.7% 33.3%	80.0% 20.0%

Fig 8: Performance Analysis of k-NN classification

IV. CONCLUSION

An novel cervical cancer growth recognition technique dependent on new order of shape features, GLCM, was displayed and assessed for 20 pap smear pictures. The outcomes exhibit the strong execution of the proposed strategy in detecting cancer growth in pictures of various resolution and quality and from different acquisition frameworks. The method has performed many automated segmentation and classification system. Cervical cancer detection is helpful for the pathologist to effectively foresee the disease organizes and incline toward treatment and treatment in like manner. The fundamental reason for proposed approach is to diminish the false negative rate by appropriate segmentation and grouping. Our methodology is performing very well on multi-cell and overlapped cells. In this procedure 95% of precision is received for KNN classifier.

REFERENCES

1. Ling Zhang, LeeLu,IsabellaNogues,Ronald M Summers,Shaoxiong Liu and jianhua Yao, DeepPap:Deep Convolutional Networks for Cervical Cell Classification , IEEE Journal of Biomedical and Health Informatics, vol 21,no 6,November 2017.
2. Dana Bazazehand RaedShubair,Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis,978-1-5090-5306-3/16/$31.00 c2016 IEEE.
3. NingGuo,Ruoh-Fang Yen,Georges EI Fakhri and QuanzhengLi,SVM Based Lung Cancer Diagnosis Using Multiple Image Features in PET/CT,978-1-4673-9862-6/15/$31.00 c2015 IEEE.
4. JonghwanHyeon,Ho-Jin Choi,Byung Doo Lee,Kap No Lee,Diagnosing Cervical Cell Images Using Pre-trained Convolutional Neural Network as Feature Extractor,978-1-5090-3015-6/17/$31.00 c2017 IEEE.
5. P.Ramachandran, N.Girija, T.Bhuvaneswari, Ph.D,Early Detection and Prevention of Cancer using Data Mining Techniques, International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014.

6. Sunny Sharma,Cervical Cancer stage prediction using Decision Tree approach of Machine Learning, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016.
7. Vikas Chaurasia1, Saurabh Pal2,A Novel Approach for Breast Cancer Detection using Data Mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014
8. Jiban K Pal ,Annals of library and Information studies,Usefulness and application of data mining in extracting information from different prespectives. Vol 58,March 2011
9. M. Durairaj, V. Ranjani ,Data Mining Applications In Healthcare Sector: A Study, international journal of scientific & technology research volume 2, issue 10, october 2013
10. Neha Sharma1 and Hari Om,Framework for early detection and prevention of oral cancer using data mining, International Journal of Advances in Engineering & Technology, Sept 2012
11. P.Ramachandran1, Dr.N.Girija2, Dr.T.Bhuvaneswari ,Cancer Spread Pattern – an Analysis using Classification and Prediction Techniques,International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June2013
12. G. Battista, C. Sassi, M. Zompatori, D. Palmarini, and R. Canini, Ground-glass opacity: Interpretation of high resolution CT findings, LaRadiolo-giaMedica , vol. 106, pp. 425–442, 2003.
13. Z. G. Yang, S. Song, and S. Talcashima, High-resolution CT analysis ofsmall lung adenocarcinoma revealed on screening helical CT, Amer. J.Roentgenol. , vol. 176, no. 6, pp. 1399–1407, 2001.
14. T. Aoki, Y. Tomoda, H. Watanabe, H. Nakata, T. Kasai, H. Hashimoto, M.Kodate, T. Osaki, and K. Yasumoto, Peripheral lung adenocarcinoma:Correlation of thin-section findings with histologic factors and survival, Radiology, vol. 220, pp. 803–809, 2001.
15. J. J. T. Owen, D. E. McLoughlin, R. K.Suniara, and E. J. Jenkinson, Therole of mesenchyme in thymus development, Current Topics Microbiol.Immunol. , vol. 251, pp. 133–137, 2000.
16. M. R. Melamed, B. J. Flehinger, M. B. Zaman, R. T. Heelan, W. A.Perchick, and N. Martini, Screening for lung cancer: Results of the memo-rialsloan-kttering study in New York, Chest, vol. 86, no. 1, pp. 44–53,1984.
17. C. V. Zwirewich, S. Vedal, R. R. Miller, and N. L. M ¨uller, Solitarypulmonary nodule: High-resolution CT and radiologic-pathologic corre-lation, Radiology, vol. 179, no. 2, pp, 469–476, 1991.
18. S. F. Huang, R. F. Chang, D. R. Chen, and W. K. Moon, Characterizationofspiculation on ultrasound lesions, IEEE Trans. Med. Imag. , vol. 23,no. 1, pp. 111–121, Jan. 2004.
19. M. Noguchi and Y. Shimosato, The development and progression ofadenocarcinoma of the lung, Cancer Treatment Res., vol. 72, pp. 131–142, 1995.
20. T. V. Colby and C. Lombard. Histiocytosis X in the lung, Human Pathol. ,vol. 14, no. 10, pp. 847–856, 1983.
21. Geert Litjens,ThijsKooi, A survey on deep learning in medical image analysis,ELSEVIER,2017
22. M. Anousouya Devi, S. Ravi, Classification of Cervical Cancer using Artificial Neural Networks, Twelfth International Multi-Conference on Information Processing-2016