

Computer Aided Detection of Cervical Cancer Using Pap Smear Images Based on Hybrid Classifier

P. Sukumar^{1*} and Dr. R. K. Gnanamurthy²

¹Assistant Professor, Department of Electronics and Communication Engineering,
Nandha Engineering College, Erode, Tamil Nadu 638052, India
Telephone: +91-9842119035, email: neuronnest@gmail.com

²Principal, Department of Electronics and Communication Engineering, SKP
Engineering College, Tiruvannamalai 606611, Tamil Nadu, India
Telephone: +91-9442246769, email: rkgnanam@yahoo.co.in

Abstract

Pap smear is a screening methodology employed in cervix cancer detection and diagnosis. The Pap smear images of cervical region are used to detect the abnormality of the cervical cells. In this paper, the computer aided automatic detection and diagnosis method for cervix cancer using Pap smear image is being presented. The hybrid classifier is used to classify the test pap smear cell image into either normal or dysplastic cell image. The abnormal cell region is detected and segmented using morphological operations. The proposed methodology is tested on the images available in publicly open access database.

Keywords: Pap smear test, Tumor segmentation, cervical cancer, Computer Aided Diagnosis, Neural networks, SVM.

Introduction

Nowadays uterus cancer disease is most common in females affecting the genital tract characterized by cervical cancer, with an increasing incidence in the last decades. The causes which lead to these cancers are the average lifespan increase in the female population above 65 years of age, and the existence of some risk factors such as unopposed estrogen replacement therapy for menopausal women, polycystic ovary disease, obesity, etc. Endometrial cancer is the second most common cancer causing death next to cervical cancer. However, it is the third most common cause of death among the female genital cancers next to ovarian and cervical cancers.

Endometrial cancer ranks as the most common gynecologic malignancy in Europe and North America, alongside ovarian cancer. More than 15% of cases of cervical cancer are found in women aged over 65 [1]. The majority of endometrial cancers

arises from glandular cells and is known as cervical adenocarcinoma [2], which seems to have become more common in the last 20 to 30 years. Surgery is the primary treatment in the majority of women with endometrial cancer. Radiotherapy is also used when adjuvant therapy is indicated. However, indication for adjuvant therapy varies from country to country. Chemotherapy and hormonal treatment are only given in clinical trial contexts or as palliation in case of an advanced stage.

The detection of cervical cancers from the Pap smear images is a challenging task in medical image processing. This can be improved in two ways. One way is by selecting suitable well defined specific features and the other is by selecting the best classifier. Many automatic and semi-automatic methods have been proposed in various times to detect various stages of cervical cancer. Many of these methods were not supported in achieving the objectives of providing measured variables which could eliminate the interpretation errors and inter-observer discrepancy.

In this paper, an automatic system based on the texture features for the classification of cervical cancer is proposed. The presented tumor detection system segments the cervical cancer from the Pap smear images into normal or dysplastic cells using Support Vector Machine (SVM) and Adaptive Neuro Fuzzy Inference System (ANFIS) classifiers. The performance of the proposed hybrid algorithm is tested and compared to other algorithms on available public image databases consisting of 40 Pap smear images. The output of classification obtained is found to be best for most of the images and the classification accuracy is 78%.

The rest part of the paper reviews some of the conventional methods of cervical cancer detection, the methodologies of proposed method and their results. The last section concludes the proposed work.

Related Works

Mariarputham and Stephen [3] presented a Nominated Texture based Cervical Cancer (NTCC) Classification System to classify the Pap smear images into any one of the seven classes. They used well-defined texture features and selected the best classifier. Seven set of texture features were extracted including the size of nucleus and cytoplasm, dynamic range and first four moments of intensities of nucleus and cytoplasm, relative displacement of nucleus within the cytoplasm, gray level co-occurrence matrix, local binary pattern histogram, tamura features and edge orientation histogram. They employed Support Vector Machine (SVM) and Neural Network (NN) classifiers for the classification process. The performance of the NTCC algorithm was tested and compared with other algorithms. The output of SVM provided a precision for normal squamous (97.38%), intermediate squamous (93.89%), mild dysplasia (87.33%), severe dysplasia (58.52%), and carcinoma in situ (84.72%) is achieved for a combination of all feature sets. With the single feature set F7, the accuracy rate of 89.35% is achieved in columnar type. Similarly, the accuracy of 84.10% for moderate dysplasia is achieved through the combination of F4 and F6 feature sets.

Yung-Fuet et al. [4] have developed an algorithm for segmenting nucleus and cytoplasm counters. This system classifies the Pap smear cells into anyone of four

different types of classes using SVM. Two experiments were conducted to validate the classification performance which showed the best performance outputs. In the first experiment, the results showed that average accuracies of 97.16% and 98.83% were obtained respectively, for differentiating four different types of cells. In the second experiment, 70% (837) of the cell images were used for training and 30% (361) for testing, achieving an accuracy of 96.12% and 98.61% for four-cluster and two-cluster classifiers, respectively.

Sobrevilla et al. [5] proposed an algorithm for nuclei detection of cytology cell. This algorithm combines color, cyto-pathologists knowledge, and fuzzy systems which show high performance and more computational speed.

Bergmeir et al. [6] have developed an algorithm used to detect cell nuclei and cytoplasm. Their algorithm used the combination of voting scheme and prior knowledge to locate the cell nuclei and elastic segmentation to determine the shape of nucleus. The noise is removed with mean-shift and median filters, edges were extracted with canny edge detection algorithm. The experimental results provided 96.69% positive predictive value (PPV), 95.63% true positive rate (TPR), and F-measure of 96.15% calculated totally for a set of 549 images.

Jusman et al. [7] have reviewed some of the cervical screening techniques, their advantages and disadvantages. The digital data of the screening techniques were used as data for the computer screening system as replaced in the expert analysis. Four stages of the computer system used in their analysis were enhancement, feature extractions, feature selection, and classification. The computer system based on cytology data and electromagnetic spectra data achieved better accuracy than other data. The accuracy in classification using Neural networks was 78.7%. The overall performances showed that cytology features and the electromagnetic spectra features provided higher accuracy of about 90%.

Harandi et al. [8] have developed a system for the detection of cytoplasm and nucleus from ThinPrep images. The geometric active contours were used as the segmentation tool. In their method, localization of cell objects were done in low resolution and boundary detection of cytoplasm and nucleus were done in high resolution.

Most of the segmentation methods discussed in this literature mainly dealt with nucleus and cytoplasm extraction, which requires higher contrast around the boundaries of nucleus. The heavily stained cervical smears, overlapping of cell images and blurred images to overexposing or underexposing of light in microscope even cause difficulties in segmentation. Thus, we propose an automatic classification of Pap smear images which overcomes the demerits of existing methods focusing primarily on classification of Pap smear cells into two classes, namely, normal and dysplastic (abnormal) cells.

Proposed Method

The proposed technique detects the cancer cells from the Pap smear image by implementation of preprocessing, Feature extraction and a hybrid classifier, combining SVM and Neural network classifiers. The preprocessing is done to remove

the noises and enhance the finer details of the image. Then, features are extracted from the preprocessed image in the training mode from a set of normal and abnormal (dysplastic) cell images in order to train the SVM and NN classifier in the training mode as shown in Fig. 1. Fig. 2 describes the testing mode of classification in which the Pap smear image is classified unto normal or dysplastic based on its extracted features. The dysplastic image thus obtained is subjected to morphological operations for nuclei segmentation and finally the performance of segmentation is evaluated.

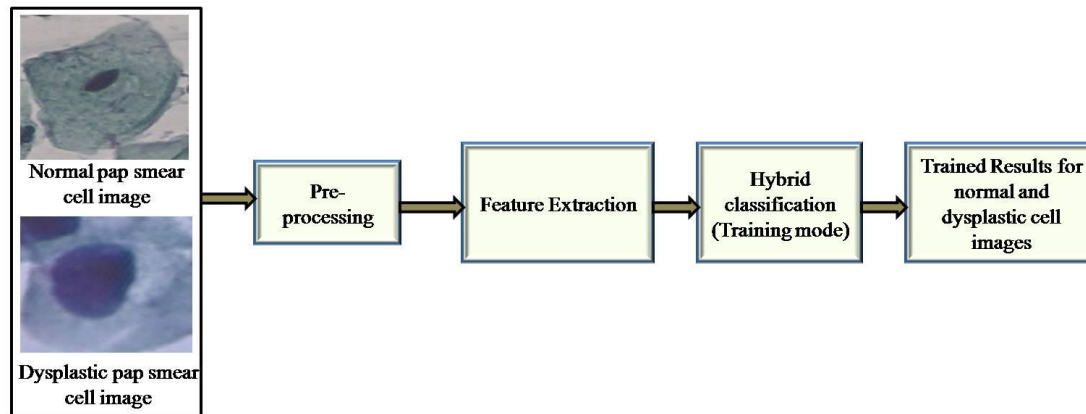


Figure 1: Hybrid Classification of Cervical Cancer In Training Mode

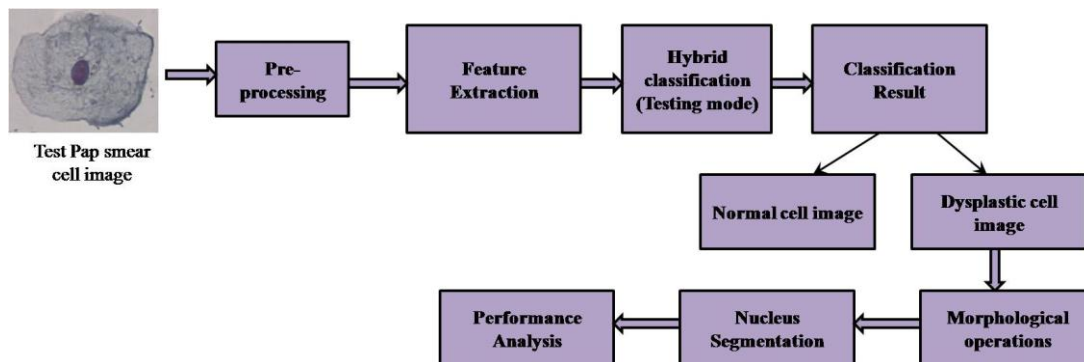


Figure 2: Hybrid classification of cervical cancer in Testing mode

Dataset

The Pap-smear images acquired from healthy & cancerous smears of patients coming from the Herlev University Hospital (Denmark)–DTU/Herlev Pap Smear Database [9] is used in this paper. These images were used by several researchers for their selected studies and the database contains papers related to these data. The old Pap Smear Database was formed in the late 90's while the improved new Pap Smear Database was formed in 2005.

We have made use of the Pap smear images from this open source database and have used the same for training and testing the classifiers. The publicly available Pap smear image dataset – DTU/Herlev is used in this paper for the performance analysis

of Pap smear cell nucleus segmentation with respect to their corresponding ground truth images.

METHODS

Pre-processing

The Pap smear cell images were acquired through a powerful microscope by the skilled cytotechnicians. All images were captured with a high resolution of $0.201\mu\text{m}/\text{pixel}$ from the public database of cervical cancer, Herlev University Hospital, Denmark [9]. In the preprocessing step, the image is resized to 256×256 pixel resolutions and RGB to grey scale conversion is implemented. The main purpose of preprocessing is that the unwanted noises are suppressed in the cervical image samples and it is enhanced for further processing.

In general, the nucleus region of the cervical cytology cell has larger distribution of darker pixel than cytoplasm. The input images are first inverted, and then the image binarization followed by morphological closing operation with structuring element of five has been performed. The cell nucleus can be approximately segmented with morphological filling operation.

Feature Extraction

In the proposed cervical cancer classification system six set of features are extracted. The features include GLCM features, LBP features, Histogram and Laws texture features, Grey level based features and wavelet features.

GLCM Features

Gray Level Co-occurrence Matrix (GLCM) is a second order statistic measurement that contains information about the positions of pixels having similar gray level values. GLCM is a $N_a\times N_a$ matrix, where N_a is the number of grey levels in the input image. The number of pixel pair repetitions are counted and updated in the GLCM matrix. GLCM identifies the texture in the given liver image, by modeling the texture as a 2-dimensional array grey level variation. This array is called Grey Level co-occurrence matrix.

Haralick has extracted many statistical features known as Haralick texture features (Haralick et al., 1973) using the GLCMs. The second-order statistics are defined as the possibility of examining a pair of grey values occurring at the endpoints of a dipole of random length placed in the image at a random location and orientation. GLCM is a statistical method that considers the spatial relationship of pixels, hence it is also known as the grey-level spatial dependence matrix. GLCM features are calculated in four directions - 0° , 45° , 90° and 145° and four distances (1,2,3,4).

To extract the features, GLCM should be a symmetric and normalized matrix. To make a matrix symmetric, transpose of GLCM is taken and added with the original GLCM. To get a normalized matrix, sum of all elements in a GLCM is calculated and each element of the matrix is divided with the obtained sum. From the normalized symmetrical GLCM, the texture features are extracted. Four properties of GLCM that

are used for our evaluation are energy, correlation, contrast and homogeneity. Consider a matrix element $P(i,j | \Delta x, \Delta y)$ which is the relative frequency of two pixels separated by pixel distance $(\Delta x, \Delta y)$. The GLCM properties within a given neighborhood with grey level intensity i and intensity j are computed as,

$$\text{Energy} = \sum p(i,j)^2 \quad (1)$$

$$\text{Homogeneity} = \frac{\sum p(i,j)}{1+|i-j|} \quad (2)$$

$$\text{Correlation} = \sum (i - \mu_i)(j - \mu_j) \frac{p(i,j)}{[\sigma_i \sigma_j]} \quad (3)$$

$$\text{Contrast} = \sum (|i - j|^2 \times p(i,j)) \quad (4)$$

Local Binary Pattern Features

The local binary pattern (LBP) operator is an efficient texture based image operator which transforms an image into an array or image of integer labels representing the small-scale appearance of the image. It works by labeling the pixels of an image by thresholding the neighborhood of each pixel and stores the result as a binary number. These labels are most commonly then used for further image analysis. The LBP operator has a computational simplicity; hence it is possible to analyze the liver CT images in real-time applications.

The LBP operator defines a two-Dimensional surface texture by 2 complementary measures, namely, grey scale contrast and local spatial patterns. The LBP operator creates labels for the pixels in the image by thresholding the 3×3 surrounding pixels with the centre value and stores the result as a binary number. In unsupervised pattern segmentation, LBP operator can be used in combination with a local contrast measure to produce higher performance.

The LBP operator is defined by, $LBP_{(J,K)}^u$ where, (J,K) represents the LBP operator used in a (J,K) neighbourhood, 'u' defines that only uniform patterns are used and labelling all other patterns with a single label. On obtaining the LBP labeled image $g_i(s, t)$, the LBP histogram H_i can be defined as,

$$H_i = \sum_{s,t} I \{ g_i(s, t) = i \} \quad i = 0, 1, \dots, n - 1 \quad (5)$$

where, 'n' is the number of LBP labels and $I\{G\}$ is 1 if G is true and 0 if G is false.

A 3×3 mask window is placed over the image and a sub-image is obtained. In the resulting 3×3 sub-image, the value of the centre pixel is compared with its neighboring pixels. If the neighboring pixel has a value greater than the centre pixel, then the neighboring pixel value is replaced by 1, or else the neighboring pixel value is replaced by 0. Finally, all the neighboring pixels will be replaced by either 0 or 1, on merging which forms an eight digit binary number. The detailed operation is explained in Fig. 3.

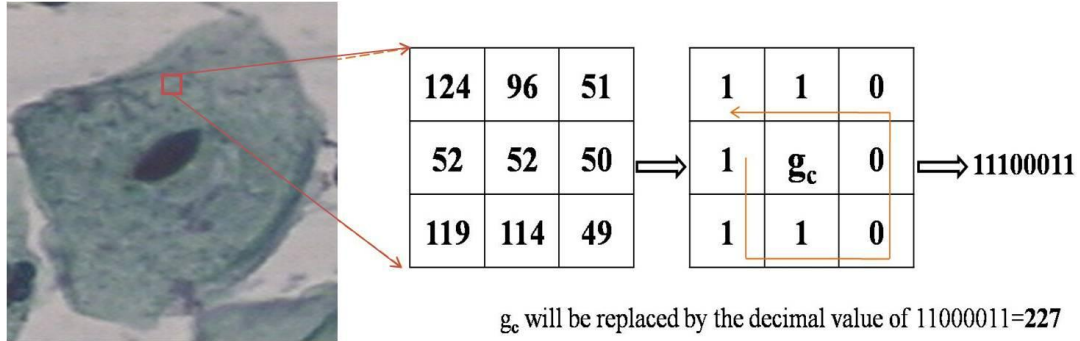


Figure 3: LBP Calculation methodology

Grey Level based Features

The grey level based features help in segmenting the Pap smear cell nuclei more easily from its background pixels based on the grey level intensity variations between a candidate pixel and its surrounding pixels. Considering a small pixel region on the given image, with the described pixel at the centre, we develop a set of grey level based descriptors for the pre-processed image. Five different feature images are extracted and considered as a feature set for the classification process. These five image sets for a candidate pixel (s, t) in a sub-image J are given as follows,

$$F_1(s, t) = I(s, t) - \min \{J\} \tag{6}$$

$$F_2(s, t) = \max \{J\} - I(s, t) \tag{7}$$

$$F_3(s, t) = \text{abs}(I(s, t)) - \text{mean} \{J\} \tag{8}$$

$$F_4(s, t) = \text{std} \{J\} \tag{9}$$

$$F_5(s, t) = I(s, t) \tag{10}$$

Histogram Features

Histogram is the most commonly used characteristic to represent the global feature composition of an image. It is invariant to translation and rotation of the images and normalizing the histogram leads to scale invariance. Histogram thresholding is widely used as a feature step for image segmentation and recognition. Its main use in this domain is to sort out the background regions from the preprocessed Pap smear image.

The x and y gradients of a grey scale image can be calculated in a number of ways. The simplest is finite differences with the [-1, 0, 1] kernel. We then use these x and y gradients to compute the orientations and magnitudes at each pixel. A pixel's orientation is denoted by a number between zero and seven. Reducing the number of different orientations to such a low number as eight, means that this can be implemented through a sequence of relationships, given by,

$$O_{x,y} = (g_y < 0).4 + (g_x < 0).2 + (|g_y| > |g_x|).1 \tag{11}$$

where, $O_{x,y}$ is the orientation at pixel (x,y), g_x and g_y are the x and y gradients at pixel (x,y). After the orientations and magnitudes of each pixel have been calculated,

the histogram for each possible sub-image in the grey scale image is estimated by summation of all orientations in that sub-image. But, only the orientations with magnitudes greater than the threshold value are included. Each of these histograms then becomes a pixel in the histogram image.

Wavelet features

A wavelet is defined as a small wave denoted by a mathematical function representing scaled and shifted copies of a finite-length waveform called the mother wavelet.

$$\chi_{a,b}(t) = \frac{1}{\sqrt{a}} \chi\left(\frac{t-b}{a}\right) \quad (12)$$

where, a and b are the scaling and shifting parameters, respectively.

A discrete wavelet transform (DWT) is based on wavelets. A DWT examines the given Pap smear image on various resolution scales and divides the image into several frequency components, i.e. a multi-resolution image is created to simultaneously view the spatial and frequency attributes. The wavelet function W is a step function denoted by a discontinuous function as:

$$W \rightarrow (a^w | s^w) \quad (13)$$

where, w is the decomposition level, a is the approximation sub-band and s is the detail sub-band. The DWT decomposes the image into four sub-bands: LL, LH, HL and HH sub-bands. (L=Low, H=High). The LL-sub-band contains an approximation of the original image while the other sub-bands contain the missing details. The output from sub-band LL, from any stage can be decomposed further. This process is called pyramid decomposition and is shown in Fig. 4.

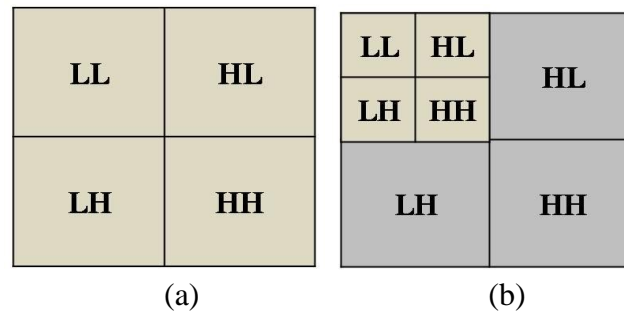


Figure 4: Pyramid decomposition using DWT at (a) Level 1, (b) Level 2

Laws Texture Features

Feature extraction is defined as the process of obtaining higher-level information from an image such as color, shape and texture. The Laws method uses filter masks to extract secondary features from natural micro-structure characteristics of the image (level, edge, spot and ripple) which are then used to segment the region of interest. Laws developed five labeled vectors which could be combined to form two dimensional convolution kernels. These masks help in extraction of individual structural components of the image when it is convolved with the image. The five vectors are: L5 = [1, 4, 6, 4, 1], E5 = [-1,-2, 0, 2, 1], S5 = [-1, 0, 2, 0,-1], R5= [1,-4,

6,-4, 1] and $W5 = [-1, 2, 0,-2, 1]$. After a series of particular convolution with selected Laws' masks, the outputs are passed to texture energy measurement (TEM) filters for the analysis of the texture property of each pixel. The energy is calculated as,

$$E = \frac{1}{MN} \sum_{x=1}^M \sum_{y=1}^N I^2(x,y) \quad (14)$$

Classification and Nucleus Segmentation

Texture classification refers to the process of grouping test samples of texture into classes, where each resulting class contains similar samples according to some similarity criterion. If the classes have not been defined a priori, the task is referred to as unsupervised classification. Alternatively, if the classes have already been defined (using training sets of sample textures) then the process is referred to as supervised classification. In this paper, all the texture classification tests reported are of the supervised type. In this paper, Support Vector Machine (SVM) and ANFIS are utilized for classification.

1) SVM Classifier:

SVMs are a set of related supervised learning methods that analyze data and used for classification and regression analysis to identify patterns. The standard SVM takes a set of input data, and estimates the input, that it is a member of which class, thereby making it a non-probabilistic binary linear classifier. The SVM training model predicts whether a new test image belongs to one category or the other based on the set of training examples.

SVM separates a set of training images two different classes, over a dimensional feature space. SVM builds the optimal separating hyper planes based on a kernel function. All images, of which feature vector lies on one side of the hyper plane, belong to class -1 and the others belong to class +1.

In other words, SVM classification involves identification of objects intimately connected to the known classes. This is called feature selection or feature extraction. SVM creates a hyper-plane between two sets of data for classification. The input data is classified such that one part of data falls on one side of the hyper-plane and another part falls on the other side. Feature selection and SVM classification together have a use even when prediction of unknown samples is not necessary.

2) ANFIS Classifier:

The adaptive neuro fuzzy inference system (ANFIS) is used to solve problems related to parameter identification. This parameter identification is done through a hybrid learning rule combining the back-propagation gradient descent and a least-squares method. ANFIS normally has 5 layers of neurons (Fig. 5) of which neurons in the same layer are of the same functional family.

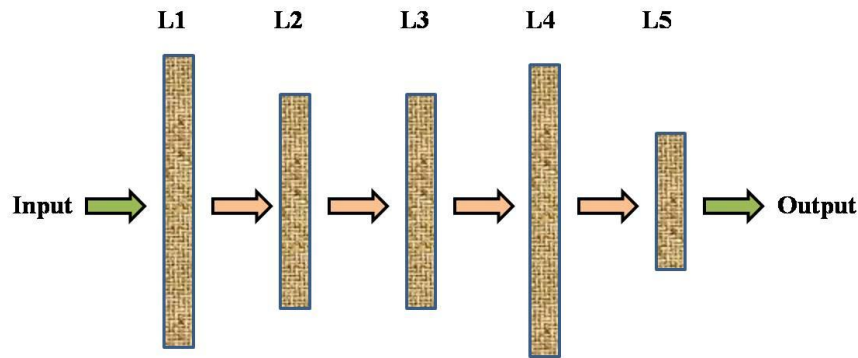


Figure 5: ANFIS Architecture

Layer 1 (L1): Each node generates the membership function denoted by,

$$\mu_A(x) = \frac{1}{1 + \left| \frac{x - c_i}{a_i} \right|^{2b_i}} \quad (15)$$

where a_i, b_i, c_i are the ground parameters to be trained.

Layer 2 (L2): In this layer, all the nodes are fixed. Here, the t-norm is used to ‘AND’ the membership grades, for example, the product given by,

$$o_{2,i} = w_i = \mu_{A_i}(x) \mu_{B_i}(y), \quad i = 1, 2 \quad (16)$$

Layer 3 (L3): The nodes in this layer calculate the ratios of the rule’s firing strength to the sum of all the rules firing strength.

Layer 4 (L4): In this layer, the nodes are adaptive and achieve the consequent of the rules:

$$o_{4,i} = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i) \quad (17)$$

The parameters in this layer (p_i, q_i, r_i) are to be determined and are referred to as the consequent parameters.

Layer 5 (L5): In this layer, a single node aggregates the overall output as the summation of all input signals, and is given by,

$$o_{5,i} = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (18)$$

3) Tumor and Nucleus Segmentation:

The tumor regions and the cell nucleus are segmented using morphological filters. The morphological operators are applied on the Pap smear cell images to segment the abnormal dysplastic regions.

- Step 1: Edges are detected using ‘sobel’ edge detector from the grey scale image.

- Step 2: Apply global thresholding technique to get ROI image.
- Step 3: Apply morphological opening function on ROI image.
- Step 4: Apply morphological eroded function on opened image.
- Step 5: Borders are cleared using ‘imclearborder’ function.
- Step 6: Final nuclei region is detected and marked in input test pap smear cell image.

Results and Discussion

The performance of proposed Pap smear cell segmentation is analyzed with the following parameters:

$$\text{Accuracy [Acc=(TP+TN)/(TP+FN+TN+FP)]} \quad (19)$$

Acc is the ratio of total well-detected and classified cell nuclei pixels. These parameters are evaluated for a set of images and listed in Table 1, where, TP denotes true positive, FP denotes false positive, FN is false negative and TN is true negative. True Positive refers to the correctly identified cell nuclei pixels, True Negative refers to the wrongly identified cell nuclei pixels, False Positive refers to the correctly identified non- cell nuclei pixels and False Negative refers to the wrongly identified non- cell nuclei pixels.

Table 1 evaluates the result of cancer cell detection accuracy for the segmentation of dysplastic cell images.

Table 1: Performance comparison

Methodology	Accuracy (%)
Proposed	99.1
Chen et al. [4]	98.16

Conclusion

In this paper, a computer aided automatic Pap smear image analysis system is proposed to analyze the cervical cancer in women. The hybrid classifier is proposed to classify the test pap smear cell image into either normal or dysplastic cell image. The texture features are extracted from the test pap smear cell image, trained and tested by hybrid classifier. The proposed system classifies the test pap smear cell image into either normal or dysplastic cell image using hybrid classifier.

References

- [1] American Cancer Society, Key statistics about Cervical Cancer. Available at: <http://www.cancer.org/cancer/cervicalcancer/detailedguide/cervical-cancer-key-statistics>, Feb 2015.

- [2] American Cancer Society, Report on Cervical Cancer Prevention and Early Detection. Available at: <http://www.cancer.org/cancer/cervicalcancer/moreinformation/cervicalcancerpreventionandearlydetection/cervical-cancer-prevention-and-early-detection-what-is-cervical-cancer>. Dec 2014.
- [3] Mariarputham, E. J., and Stephen, A., 2015, "Nominated Texture based Cervical Cancer Classification," *Computational and Mathematical Methods in Medicine*, 2015(586928), pp. 1–10.
- [4] Chen, Y. F., Huang, P. C., Lin, K. C., Lin, H. H., Wang, L. E., Cheng, C. C., Chen, T. P., Chan, Y. K., Chiang, J. Y., 2014, "Semi-automatic segmentation and classification of Pap smear cells," *IEEE J Biomed Health Inform.*, 18(1), pp. 94–108.
- [5] Sobrevilla, P., Montseny, E., Vaschetto, F., and Lerma, E., 2010, "Fuzzy-based analysis of microscopic color cervical pap smear images: Nuclei detection," *Int. J. Comput. Intell. Appl.*, 9(3), pp. 187–206.
- [6] Bergmeir, C., Garcia-Silvente, M., and Benitez, J.M., 2012, "Segmentation of cervical cell nuclei in high-resolution microscopic images: A new algorithm and a web-based software framework," *Comput. Methods Prog. Biomed.*, 107(3), pp. 497–512.
- [7] Jusman, Y., Cheok Ng, S., and Osman, N. A. A., 2014, "Intelligent Screening Systems for Cervical Cancer," *Scientific World Journal*, 2014(810368), pp. 1–15.
- [8] Harandi, N. M., Sadri, S., Moghaddam, N. A., and Amirfattahi, R., 2010, "An automated method for segmentation of epithelial cervical cell images of ThinPrep," *Journal of Medical Systems*, 34(6), pp.1043–1058.
- [9] DTU/Herlev Pap smear Database, Herlev University Hospital (Denmark), MDE Lab [Online]. Available at: <http://labs.fme.aegean.gr/decision/downloads>.