ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Computer Aided Screening of Cervical Cancer Using Random Forest Classifier.

P Sukumar[1]*, and RK Gnanamurthy[2].

[1]Department of Electronics and Communication Engineering, Nandha Engineering College, Erode, Tamil Nadu 638052, India.
[2]Department of Electronics and Communication Engineering, SKP Engineering College, Tiruvannamalai 606611, Tamil Nadu, India.

ABSTRACT

Cervical cancer is a pathological disorder of the human that affects around 50 million people worldwide. The early detection and diagnosis of cervical cancer is important for the women to prevent it in an earlier stage. Conventional approaches used abnormal lesion size and shape to detect the cancer area, which provides low accuracy for surgery. In this paper, computer aided automatic cervical cancer detection methodology is proposed to overcome the limitations of the conventional methods. In preprocessing stage, complex wavelet transform is used to convert the time domain image into multi resolution image and further features are extracted from this transformed wavelet coefficients.Then,the random forest classifier is used to train and test the cervical cancer.

Keywords: Cervical Cancer, Screening, Cervigram, lesions,complex wavelet, multi resolution;

**Corresponding author*

INTRODUCTION

The cervix is located in the lower part of the uterus also called uterine cervix, it connects the body of the uterus by the cervix part called endocervix to the birth canal by the part named exocervix. Cells covering the cervix are referred to as squamous cells and the glandular cells [1]. More than 15% of cases of cervical cancer are found in women aged over 65 [2]. 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.

Cervical cancer is one of the leading causes of death for middle-aged women in developing countries even though the disease is almost completely preventable at the earlier stage if precancerous lesions are detected and treated on time. There are several methods for control and prevention of cervical cancer which includes conventional cytology like Pap smear test, liquid-based cytology, human papillomavirus (HPV) screening and vaccination against HPV. Cytology-based and HPV screening methods are hard to implement in developing countries and hence the use of visual screening by use of acetic acid (VIA) test is mostly preferred to identify the cervical cancer.

According to the World Health Organization [3], cervical cancer is said to be the world's second deadliest cancer and about 493,243 women are diagnosed with the disease and about 273,505 deaths occurring per year. Cervical cancer is also the most common cause of female genital cancers and female cancer deaths worldwide [4].

Cervical cancer develops in the cells around the cervix in the beginning stage. Pre-cancerous cells which are described as cervical intraepithelial neoplasia (CIN), squamous intraepithelial lesion (SIL) and dysplasia. The pre-cancerous cells cancer can fully grow into cancer. There are two main forms of cervical cancer namely squamous cell carcinoma and adenocarcinoma, of these types 80% to 90% of the cervical cancers are due to the squamous cell carcinoma which begin where the exocervix joins the endocervix.

Section 2 presents the related works on the detection methodologies of cervical cancer. Section 3 presents the dataset used and methodologies used to detect and diagnose the cervical cancer from cervix images using Random Forest classifier. Section 4 presents the results and discussion and Section 5 concludes the work.

RELATED WORKS

Park et al. [5] presented a domain-specific automated image analysis framework for the detection of cancerous lesions of the cervix using conditional random fields. A novel window-based performance assessment scheme for 2D image analysis was proposed. Image regions corresponding to different tissue types are indentified for the extraction of domain-specific anatomical features. The method was evaluated using clinical data from 48 patients and the results produced a standard deviation of 0.018 in the Dice value, compared to 0.022 for the colposcopy annotations.

Fan et al. [6] proposed an edge detection algorithm for alternative edge characteristic is that the colors (or grey level) of the neighboring pixels are considerably dissimilar, although their brightness values are same. Therefore, the study claimed that both the brightness and changes in the grey level between neighboring pixels should be exploited for more efficient edge extraction.

Song et al. [7] proposed a multimodal entity coreference for diagnosis of cervical Dysplasia. A data-driven computer algorithm was developed for interpreting cervical images based on color and texture. This comprehensive algorithmic framework based on Multimodal Entity Coreference was used for combining various tests to perform disease Classification and diagnosis. This method failed to detect the edge boundaries of the cervical cancer region. However, the authors achieved sensitivity of 83.21% and specificity of 94.79% for the detection of cervical dysplasia.

Many automatic [8–10] and semiautomatic [11–12] methods have been proposed to detect nuclear contours of cervical cells on the Pap smear images and to discriminate the normal from abnormal dysplasia cells. Chou et al. [13] have proposed a method used hierarchical multiple classifier scheme. This method used graph-theoretic clustering algorithm to group the training data, component classifiers as the inputs to a super-classifier, and sub class labeling is used to improve the classification accuracy.

Holmquist et al. [14] developed a binary classification method to distinguish between normal and abnormal cells. The dual wavelength method was used for the automatic isolation of nucleus from cytoplasm. The classification procedures were done based on the extraction of density-oriented, shape-oriented and texture-oriented parameters.

PROPOSED METHOD

The flow of proposed methodology in training mode and classification mode are depicted in Figure 1 and Figure 2. The cervical images are preprocessed at the initial stage and then enhanced. The preprocessing is done to remove the noises and enhance the finer details of the image.

After the preprocessing step, the enhanced cervical image is applied with complex wavelet transform and the features are extracted in the training mode from a set of normal and abnormal cervix images. These features are used to train the Random forest classifier in the training mode. The same processing steps are then implemented for any test image in the classification mode. The results of classification are then grouped into abnormal and normal images and finally compared with their corresponding ground truth images for performance evaluation.

Dataset

The digital cervix images obtained from women who participated in National Cancer Institute’s (NCI) Guanacaste study [15] are used in our research. The women can be categorized as follows: patients with invasive cancer, patients without cervical lesion at enrolment but later developed disease at follow-up and healthy women who never developed any pathological changes in the cervix. The resolution of these images is 2891×1973 pixels. The various cervical images in Guanacaste database of different patients at various stages are used in the training phase.

The proposed cervical disease classification system is evaluated on 280 randomly selected participants from this Guanacaste database. An unbalanced number of patient cases are chosen for the four categories because only 10 cancer cases are available in the entire Guanacaste dataset. However, since binary classification is carried out, i.e., classifying patient cases into one of the following two categories: <CIN2 (mild cases) and CIN 2/3+ (severe cases), such that there exists equal number of patient cases in these two classes: 140 cases in <CIN2, and 140 cases in CIN 2/3+.

Pre-processing

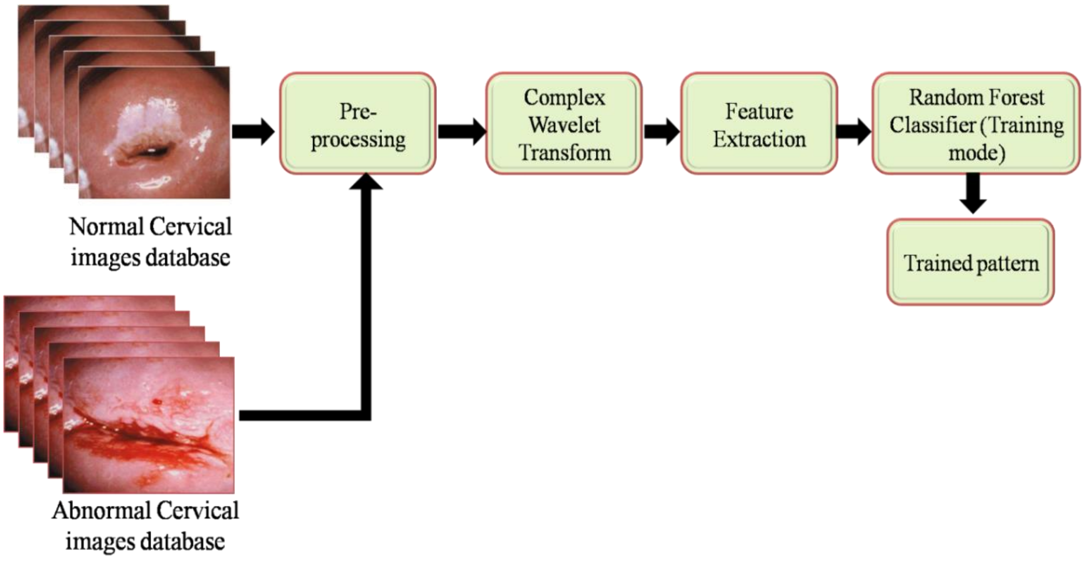

Normal Cervical images database
Abnormal Cervical images database
Pre-processing
Complex Wavelet Transform
Feature Extraction
Random Forest Classifier (Training mode)
Trained pattern

Figure 1: Proposed flow of cervical cancer detection system in training mode.

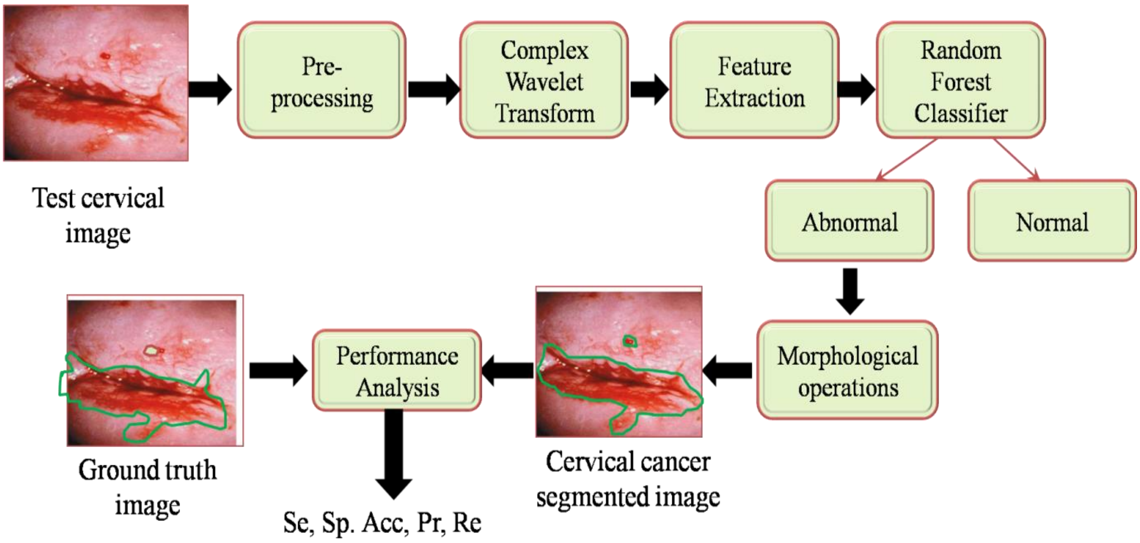


Figure 2: Proposed flow of cervical cancer detection system in testing mode.

In image analysis, Pre-processing is the primary step which performs image enhancement and reduces noise, thus, enhancing the image quality. The pre-processing phase is necessary for the extraction of the background, in order to improve the fine details for better analysis. Firstly, the original cervical image is converted to gray scale image.

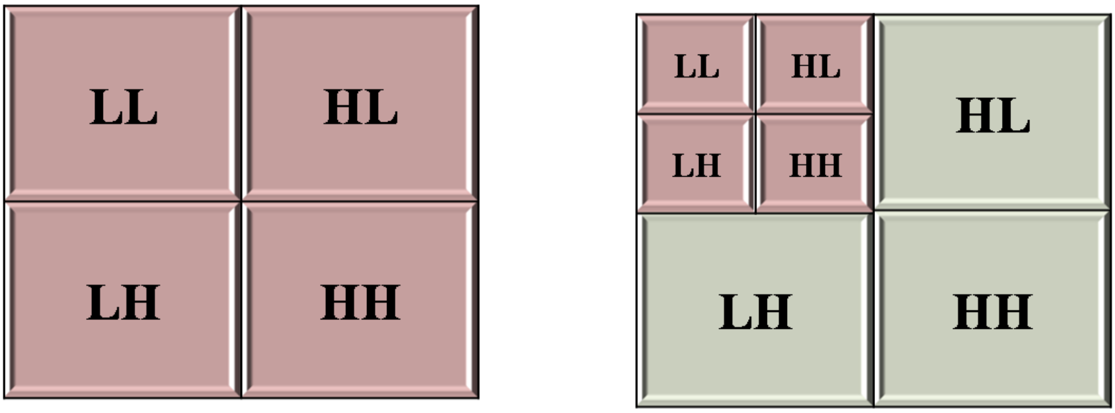
Then it is resized to 256×256 pixels suitable for further processing. The color image of cervix is converted to gray scale image by using a large matrix whose entries are numerical values between 0 and 255, where 0 corresponds to black pixels and 255 corresponds to white pixels.

Complex Wavelet Transform

Wavelet Transform is a texture related feature which is formed by the construction of wavelets from the cervical image. Wavelets are small waves and are mathematical functions that represent scaled and shifted copies of a finite-length waveform called the mother wavelet. The wavelet can be defined as,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \qquad (1)$$

where, ‘a’ is the scaling parameter and b is the shifting parameter.



(a) (b)

Figure 3: Pyramid Decomposition using discrete wavelet transform: (a) Decomposition at Level 1, (b) Decomposition at Level 2.

The wavelet transform is applied to each row and secondly to each column of the resulting image of the first operation. The resulting image is decomposed into four sub-bands: LL, HL, LH, 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 LL-sub-band output from any stage can be decomposed further. Figure 3 shows the result of pyramid decomposition.

FEATURE EXTRACTION

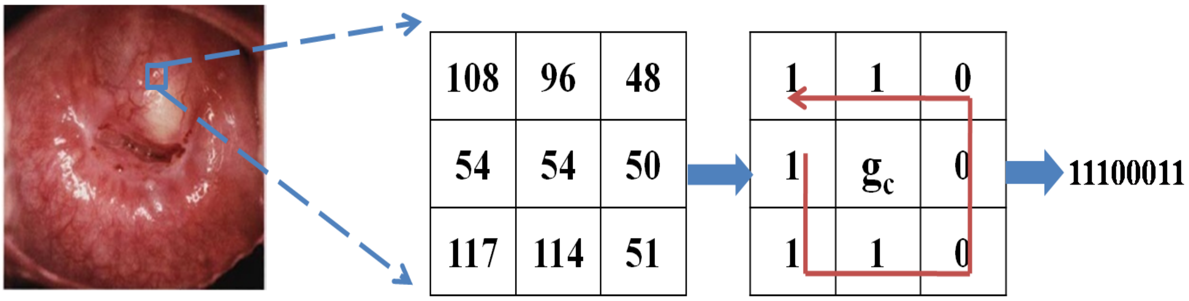
Local Binary Pattern (LBP) Features

The Local Binary Pattern (LBP) operator is an image operator discovered by Ojala et al. in 1996 which is used to transform an image into an array or image of integer labels which describes the small-scale appearance of the image. The actual local binary pattern operator works on a 3×3 pixel block of an image, i.e. sub-image. The pixels in the sub-image are thresholded by its centre pixel value, multiplied by powers of two and then added up to get a new label (value) for the central pixel.

In a 3×3 sub-image, the neighborhood consists of 8 pixels, thus, a total of 2^8 = 256 different labels can be obtained depending on the relative gray values of the centre pixel and its neighbors. The generic local binary pattern operator is derived from the joint distribution. As in the case of basic LBP, it is obtained by summing the thresholded differences weighted by powers of two. The $LBP_{P,R}$ operator is defined as

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{p=1} s(g_p - g_c)2^p \qquad (2)$$

A sub-image is obtained by placing a 3×3 mask window over the original image. 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 equal to or 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, an eight digit binary number is formed after all the surrounding pixels are replaced. This is then converted to its corresponding decimal value and is replaced with the centre pixel. The detailed operation is explained in Figure 4.



108	96	48
54	54	50
117	114	51

1	1	0
1	g_c	0
1	1	0

g_c will be replaced by the decimal value of 11000011=**227**

Figure 4: LBP Computation procedure.

Local Ternary Pattern (LTP)

Local binary pattern (LBP) is extended to a three-valued code called the Local Ternary Pattern (LTP), in which gray values in the zone of width $\pm t$ around r_c are quantized to zero, those above $(r_c + t)$are quantized to +1, and those below $(r_c - t)$ are quantized to –1, i.e., the indicator is replaced with three-valued function, as shown in Figure 5. The reason for choosing LTP as a feature selection method is that it provides ternary values where LBP provides only binary values.

$$f(k,g_c,t)=\left|\begin{array}{ll}+1, & k\geq r_c+t\\ 0, & |k-r_c|<t\\ -1, & k\geq r_c-t\end{array}\right|k=r_p \qquad (3)$$

The standard local ternary pattern (LTP) encodes the relationship between the referenced pixel and its surrounding neighbors by computing their gray-level differences.

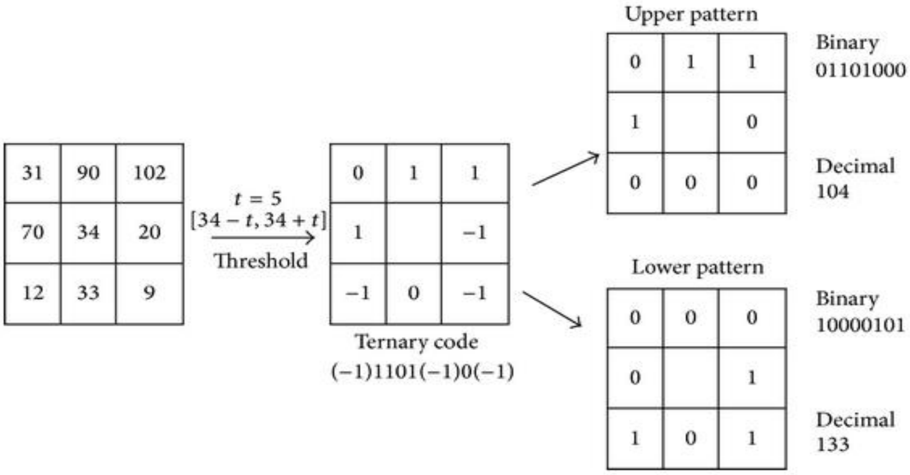


Figure 5: An example for computation of LTP.

In Figure 5, the center pixel for 3×3 sub image is 34. The extracted local ternary patterns are 104 and 133 for the center pixel 34 in this example. Next, 3×3 mask is moved over all the pixels of the entire image and generates the corresponding LTP features.

Local Derivative Pattern (LDP)

LBP is considered as the non-directional first-order local pattern operator and it is extended to higher orders called the Local Derivative Pattern (LDP). The LDP contains more detailed discriminative features as compared with the LBP. To calculate the k^{th}-order LDP, the $(\mathrm{k}-1)^{\mathrm{th}}$-order derivatives are calculated along different directions (φ), i.e. 0°, 45°, 90°, and 135°, and is denoted as,

$$\mathrm{LDR}_{\varphi}^{(k-1)}(\mathrm{r}_c)\big|_{\varphi} = 0^\circ, 45^\circ, 90^\circ, 135^\circ \qquad (4)$$

Finally, the k^{th}-order LDP is calculated for a center pixel over the eight neighboring pixels (P) as,

$$\mathrm{LDR}_{\varphi}^{k}(\mathrm{r}_c) = \sum_{\mathrm{p}=1}^{\mathrm{P}} 2^{(\mathrm{p}-1)} \times \mathrm{f}_2(\mathrm{I}_{\varphi}^{(k-1)}(\mathrm{r}_c), \mathrm{I}_{\varphi}^{(k-1)}(\mathrm{r}_p))\Big|_{\mathrm{P}} = 8 \qquad (5)$$

$$\text{Where, } f_2(x,y) = \begin{cases}1, & for\ x,y \geq 0\\ 0, & else\end{cases}$$

Random Forest Classification

The Random Forest classifier is implemented with weighted voting to classify the normal and abnormal cervical image. Consider the forest to be composed of ‘N’ number of random trees initially and the complete training set be denoted by S. We select *s* number of training feature vectors from S with replacement for each tree. The training feature vector k ∈ s, is provided with ‘f’ number of features available from F(k) with replacement. Each tree is grown using ‘f’ features and each split point was selected based on a feature f∈f thereby maximizing the information gain.

The weighted voting mechanism has been implemented in the forest to classify the test data of cervix. The weight mechanism prevents the trees with poor classification performance during training phase. During classification, the class label of the test data is determined by the weighted voting from the trees. Finally, the cervical images are classified based on the features trained by the random forest classifier. After the proper detection of abnormal cervix, the morphological operations are performed to segment the cancerous regions in the cervical image.

Morphological Operations for Cancer Region Segmentation

The morphological operations are applied over the abnormal images of cervix after detection by the Random Forest classifier. The functions of morphological operators are to separate the tumor affected portion from the cervical image and mark the region for further tumor diagnosis. The morphological operations are applied on the gray scale cervical image to segment the abnormal regions. Erosion and dilation are the two elementary operations in mathematical morphology. An aggregation of these two characterizes the rest of the operations. The symbols $\oplus$, $\ominus$, $\circ$, and $\bullet$ denote the four fundamental binary morphological operations: dilation, erosion, opening, and closing, respectively. A function R(x, y) or R denotes the image, and the function H(x, y), or H denotes the structuring element. The four operations are defined as follows:

Dilation:

$$(R \oplus H)\,(x,y) \;=\; \overset{sup}{{}_{(r,a)\in H}} \{x - r,\; y - s) + H(r,s)\} \tag{6}$$

Erosion:

$$(R \ominus H)\,(x,y) \;=\; \overset{inf}{{}_{(r,a)\in H}} \{x + r,\; y + s) - H(r,s)\} \tag{7}$$

Opening:

$$R \circ H = (R \ominus H) \oplus H \tag{8}$$

Closing:

$$R \bullet H = (R \oplus H) \ominus H \tag{9}$$

Where, *sup*{} and *inf*{} refers to the supremum and infirmum operations, respectively. Erosion and Dilation are combined to form a powerful operator called Opening, by which objects that are adjacent are spaced and objects that are adjoined are detached and the holes within the objects are enlarged. Finally, the tumor portions in the cervical image become visible due to its high intensity.

RESULTS AND DISCUSSION

Numerous simulations were performed and the results are given in this section. The original image used for our experimentation is shown in Figure 6a. The proposed method segments the cancer regions from the feature sets and the abnormal regions are marked in the cervical image in green color. The features such as LBP, LTP and LDP are extracted from the preprocessed cervical image and then processed with our proposed cervical cancer detection algorithm. Finally, the Random forest classifier and morphological operators segment the tumor regions and the resultant tumor segmented image is shown in Figure 6c.

The performance of cervical cancer region segmentation is analyzed with the following parameter:

Accuracy [Acc=(TP+TN)/(TP+FN+TN+FP)] (10)

Acc is the ratio of total well-detected and classified cervical cancer region. Table 1 evaluates the result of performance parameter for the segmentation of cancer regions of cervix. The proposed method is compared with the segmented results of the conventional method and the conventional results are shown in Figure 6d.

Table 1: Performance measure showing Accuracy of proposed method.

Test Images	Accuracy (%)
Image 1	92.8
Image 2	96.4
Image 3	93.9
Image 4	98.7
Average	**95.4**

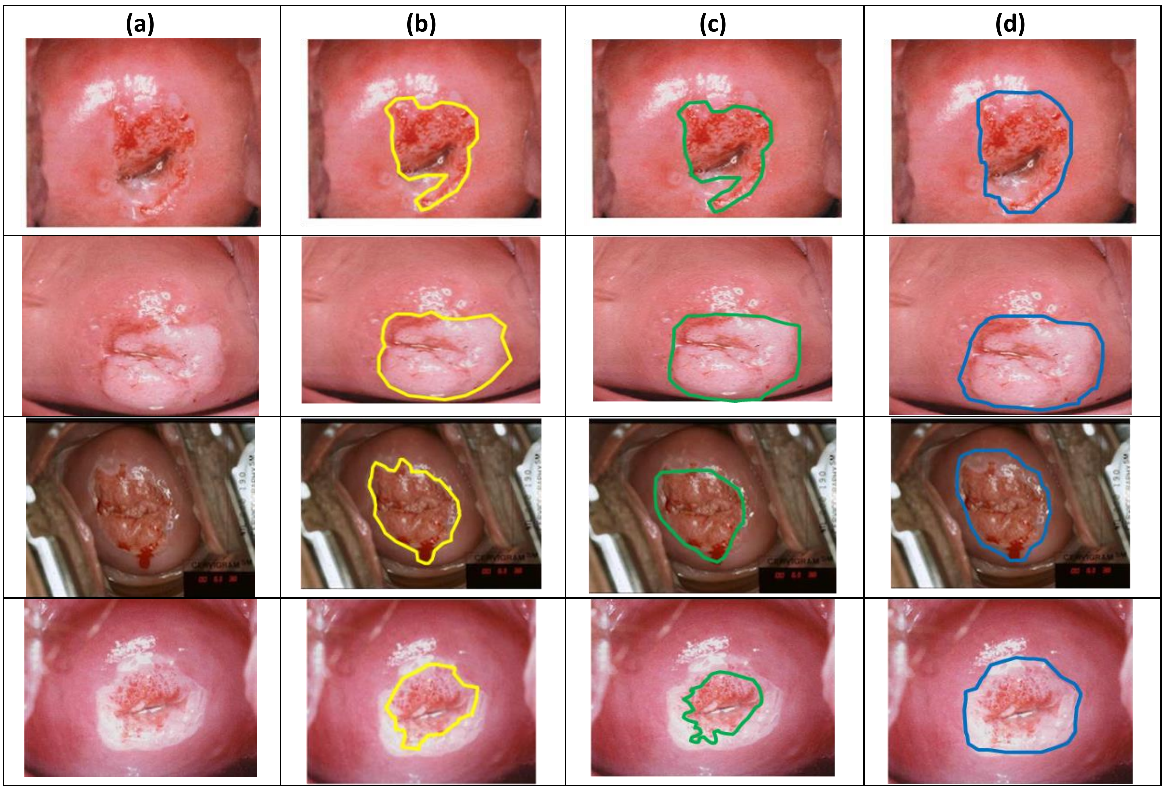


Figure 6: Column a Source Images, **Column b** Ground truth images of corresponding source images obtained from radiologist (yellow encircled region), **Column c** Cervical cancer segmented image by proposed method (green encircled region), **Column d** Cervical cancer segmented image by conventional method [18] (blue encircled region).

The proposed method is compared with other conventional methods of cervical cancer detection. The parameters used for comparison of performance include Sensitivity, Specificity and Accuracy. The various methods used different classifier methodologies and produced lower accuracies of classification, whereas the proposed system achieved a higher accuracy compared to the existing methods and is tabulated in Table 2.

Table 2: Performance comparisons of proposed method.

Methodology	Accuracy (%)
Proposed system	**95.4**
Kim et al. [16]	76
Chang et al. [17]	83
Davey et al. [18]	93.8

CONCLUSION

In this paper, computer aided automatic detection and segmentation of cancer region in cervical images is proposed using random forest classifier. The complex wavelet transform is applied on the cervical gray scale image in order to convert the spatial domain image into multi resolution image. Then, the texture features are extracted from this transformed image and the features are fed to the training mode of the random forest classifier. Finally, the test image features are fed in to testing mode of this classifier to detect and segment the cancer region in the cervical images.

REFERENCES

[1] http://www.cancer.org/cancer/cervicalcancer/detailedguide/cervical-cancer-what-is-cervical-cancer.

[2] http://www.cancer.org/cancer/cervicalcancer/detailedguide/cervical-cancer-key-statistics.

[3] http://whqlibdoc.who.int/publications/002/9241545720.pdf

[4] World Health Organization. (2004) Global burden of disease report: Causes of death in 2004. Global Burden of Disease Report. Geneva: World Health Organization.

[5] Park SY, Sargent D, Richard Lieberman, Ulf Gustafsson. IEEE Transactions on Medical Imaging 2011; 30(3): 867–878.

[6] Fan JP, Yau DKY, Elmagarmid AK, Aref WG. IEEE Transaction on Image Processing 2001; 10: 1454–1466.

[7] Song D, Edward Kim, Xiaolei Huang, Joseph Patruno, Héctor Muñoz-Avila, Jeff Heflin, Rodney Long L, Sameer Antani. IEEE Transactions on Medical Imaging 2015; 34(1): 229–235.

[8] Plissiti ME, Nikou C, Charchanti A. IEEE Trans Inf Technol Biomed 2011; 15: 233–241.

[9] Sobrevilla P, Montseny E, Vaschetto F, Lerma E. Comput Intell Appl 2010; 9: 187–206.

[10] Harandi NM, Sadri S, Moghaddam NA, Amirfattahi R. J Med Syst 2010; 34: 1043–1058.

[11] Bergmeir C, Garcia-Silvente M, Benitez JM. Comput. Methods Prog Biomed 2012; 107: 497–512.

[12] Sulaimana SN, Isab NAM, Othmanc NH. Int. J. Knowledge-based Intell Eng Syst 2011; 15: 131–143.

[13] Chou YY, Shapiro LG. Pattern Analysis and Applications 2003; 6(2): 150–168.

[14] Holmquist J, Bengtsson E, Eriksson O, Nordin B, Stenkvist B. The J Histochem Cytochem 1978; 26(11): 1000-1017.

[15] Herrero R, Schiffman M, Bratti C, Hildesheim A, Balmaceda I, Sherman M. Rev Panam Salud Publications 1997; 1: 362–375.

[16] Kim E, Huang X. A data driven approach to cervigram image analysis and classification. Color Medical Image Analysis, ser. Lecture Notes in Comput. Vis. Biomechan., M. E. Celebi and G. Schaefer, Eds. Amsterdam, The Netherlands: Springer 2013; 6: 1–13.

[17] Chang S, Mirabal Y, Atkinson E, Cox D, Malpica A, Follen M, Richards-Kortum R. J Lower Genital Tract Dis 2005; 10(2): 24–31.

[18] Davey E, Assuncao J, Irwig L, Macaskill P, Chan SF, Richards A, Farnsworth A. Br Med J 2007; 335(7609): 31.