Detection of Cervical Cancer using GLCM and Support Vector Machines

Shalini Nehra, Jagdish Lal Raheja
Control and Automation Group,
CSIR-CEERI, Pilani, INDIA
shalininehra17@gmail.com, jagdish@ceeri.res.in

Kaustubh Butte, Ameya Zope
BITS Pilani,
KK Birla Goa Campus, INDIA
kaustubhbutte@gmail.com, zopeameya@gmail.com

***Abstract*— Early detection of cancer can lead to a higher likelihood of survival, lower costs of care which would further result in lowering death rates and disability due to cancer. Introduced in March 1924, colposcopy was used to detect the cause of abnormal looking cervix and hence provide appropriate treatment. Being prone to human errors, colposcopy can lead to misleading results for detection of cervical cancer. This paper aims at presenting a convenient and automated method of detection of cervical cancer and classifying images as cancerous or non-cancerous. After applying scaling on the images obtained via colposcopy, the Gray Level Co-occurrence Matrix (GLCM) was constructed. The Haralick features extracted from the above-constructed matrix were hence used to classify images into cancerous and non-cancerous using Support Vector Machines (SVM).**

Keywords*— Gray Level Co-Occurrence Matrix (GLCM), Haralick texture features, Support Vector Machine (SVM).

I. INTRODUCTION

According to the World Health Organization (WHO), cervical cancer is the fourth most frequent cancer in women with an estimated 530,000 new cases in 2012 representing 7.9% of all female cancers. Approximately 90% of the 270,000 recorded deaths resulting from cervical cancer in 2015 occurred in low and middle-income countries. According to Cancer Research UK [1] on a survey conducted in 2012 more than 265,000 women died from cervical cancer across the world. This signifies the need for cheaper and earlier detection of cervical cancer. Particularly in countries where cervical cancer screening programs are not available, diagnosing cervical cancer at an early stage and providing access to effective treatment can significantly improve the likelihood of survival. Owing to the fact that an increasing number of individuals are gaining access to better and more advanced healthcare and undergoing Pap smear test, and there is a shortage of skilled and experienced pathologists, on the other hand, reviewing Pap smear slides has become quite a time consuming which ultimately leads to an increase of workload on the pathologists and can also load to fatigue. This can be a potential cause of inaccuracies in diagnosis. In order to solve this problem, an automated detection system of cervical cancer has been proposed [2].

In a research conducted by B.Ashok and Dr.P.Aruna [3] feature selection methods for the diagnosis of cervical cancer using SVM classifier have been specified. Although the accuracy in the above-mentioned research reaches 98.5%, their method requires the use of a microscope with an inbuilt camera attached in order to take pictures and feed to the prediction algorithm. The method presented in this research paper requires only the use of a camera, which in today's world is easily accessible. D. Kashyap et. al. [4] achieved an accuracy of 95% using SVM classifier and polynomial kernel on images of cancerous and non-cancerous images of cells. This again necessitates the use of a microscope with an inbuilt camera. The paper by Soumya MK et. al. [5] proposes to classify cervical cancer images into different stages based on the treatment volume that the particular patient requires but for that it is required for them to know that the patient is affected with cervical cancer which is proposed in our paper.

The paper by Mustafa et. al. [6] states that though Pap test is the most popular and effective test for cervical cancer, Pap test does not always produce a good diagnostic performance. The paper by P. Mohaniah et. al. [7] mentions that as the size (dimension) of the image increases, the value of texture features extracted from them also increases. So, this paper proposes an optimal size of the image i.e. 128x128 for feature extraction for better resolution and minimum loss of generality. S. Kaaviya et. al. [8] proposes a new approach to identify abnormal cervical cells using the area of the nucleus as the feature for classification. K.Pradeep Chandran, et al.[9] presents a segmentation method, a spatial fuzzy clustering algorithm, for segmenting Magnetic Resonance images to detect the Cancer in its early stages anatomical structures. Here a Probabilistic Neural Network with radial basis function is employed to implement an automated Tumor classification. Priyanka K Malli and Dr Suvarna Nandyal [10] proposed an automated, comprehensive machine learning for the detection of cervical cancer. Using the colour and shape features of nucleus and cytoplasm of the segmented unit of the cervix cell, they propose to train a k-NN and an Artificial Neural Network (ANN). The above-mentioned approach has shown an accuracy of 88.04% for KNN and 54% for ANN. Our method gives an accuracy of 96.67% in classifying images into cancerous and non-cancerous. Chankong et. al. [11] propose a method that utilizes a set of simple features extracted from the two- dimensional Fourier transform of the cell images in order to avoid the problem of cell and nucleus segmentation. The features used are calculated based on the mean, variance, and entropy obtained from the frequency components along the circle of radius (r) centred at the centre of the spectrum and the frequency components along the radial line at an angle θ. The above-mentioned approach achieves an accuracy of 92% classification rate on a set of 276 cervical single cell images containing 138 normal cells and 138 abnormal cells. Setu Garg et. al. [12] have proposed the use of edge detection and hybrid segmentation to extract a tumour infected area. Sajeena T A et. al. [13] propose a method of cervical cancer detection using Radiation Gradient Vector Flow segmentation and SVM and artificial neural networks. H. G. Acosta-Mesa et al [14] proposed a method to use the Bayesian learning for training model. Zhao M et al [15] proposed a method for cervical cancer screening algorithm

978-1-5386-7050-7/18/$31.00 ©2018 IEEE

based on WSCCI method. The experimental result shows 98.90% accuracy, but it works only on a block of images. [16] proposes an SVM-based feature screening method which to multispectral Pap smear image classification for cervical cancer detection. Experiments Results show significant improvements in pixel-level classification accuracy but this algorithm is tested for a dataset containing 40 images.

II. Approach to Cervical Cancer Detection

Algorithm flow of cervical cancer detection method has been shown in Fig 1. It includes seven steps. The presented methodology of automated classification of colposcopy cervix images starts with the collection of colposcopy images collected from freely available images on the image was constructed. Once the GLCMs were available we extracted the 13 Haralick features [17] for each GLCM. The SVM classifier was then trained using the 13 features extracted for each image in the training dataset. After training the SVM Model [18], the images present in the testing dataset were classified by using the above mentioned SVM Model. After predicting the classes on the test dataset, performance analysis was done. The dataset thus obtained was divided into 2 parts-

1. Training Dataset = 70%of original dataset.
2. Testing Dataset = 30%of original datasets

The algorithm works as follows: -

1. The cervical cancer colposcopy test images were classified and labelled accordingly.
2. After segregating the images, the images were read by the program.
3. The basic pre-requisite for calculating the GLCM is the presence of only one channelled images as input. This was done by converting the 3- channelled RGB images to the 1- channelled grayscale image. The pixel intensities of the images thus obtained was scaled down from 0-255 to 0-7 uniformly.
4. The GLCM was calculated using ϴ=0 and radius=1.
5. All the second order texture features mentioned by Haralick et. Al [19] were extracted using the GLCM obtained in the above step.
6. SVM was trained and applied to the dataset.

A labeled dataset of images was obtained from nearby hospitals. It consisted of 1012 images with 528 cancerous and 484 non-cancerous images. The images in the dataset were labelled as cancerous or non-cancerous. These images were given as input to the program which first performed scaling of each of the input images. All the images were then converted to grayscale in order to calculate the GLCM (Gray Level Co-occurrence Matrix). The transformation of a Non-cancerous image from an input Image Fig 2(a) to the grayscale image is shown in Fig 2(b) and transformation of cancerous image RGB to grayscale is shown in Fig. 3.

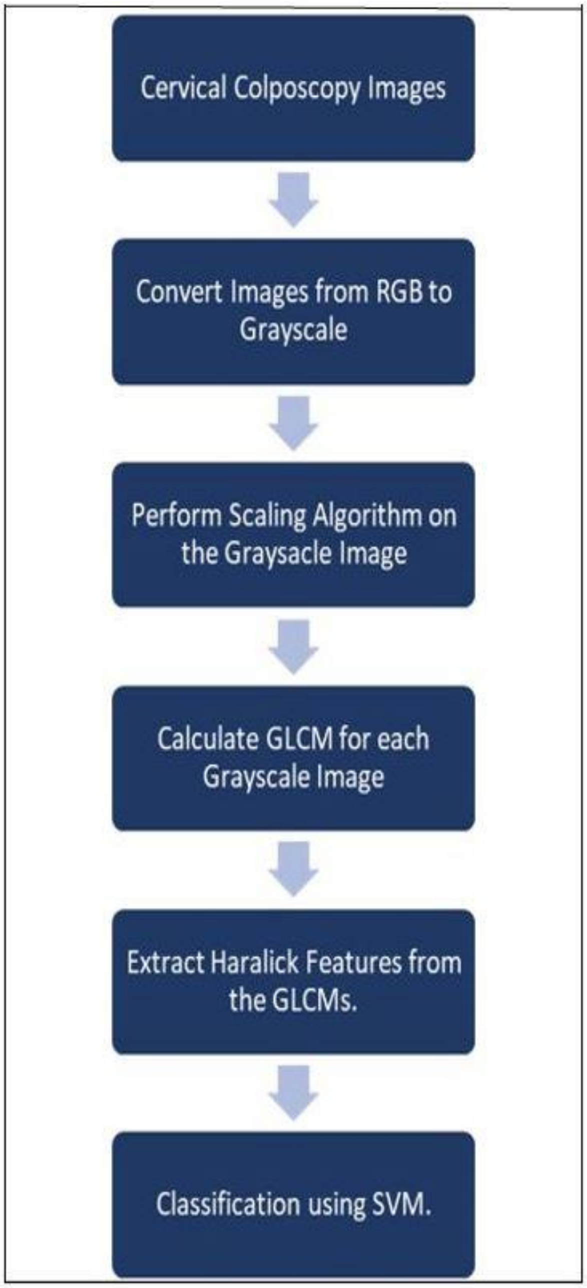
Cervical Colposcopy Images
Convert Images from RGB to Grayscale
Perform Scaling Algorithm on the Graysacle Image
Calculate GLCM for each Grayscale Image
Extract Haralick Features from the GLCMs.
Classification using SVM.

Fig. 1. Flowchart of Methodology

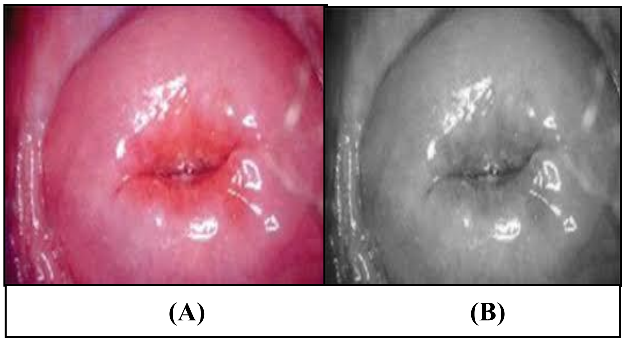

Fig. 2 Non-cancerous Image (A) RGB (B) Grayscale

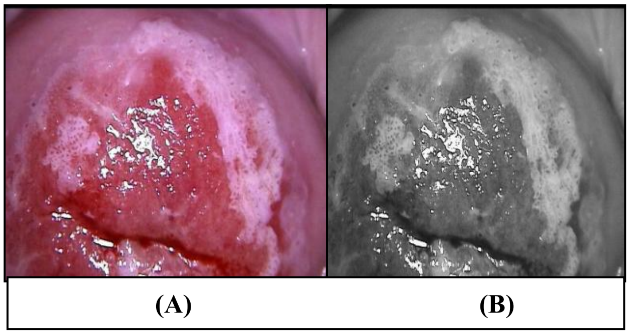

Fig. 3 Cancerous Image (A) RGB (B) Grayscale

The pixel intensities present in the grayscale image were integers ranging from 0-255 (including both values). We then created a mapping from the domain [0,255] to the codomain [0-7] which distributed the range of the domain uniformly to the range of the co-domain. After applying the above-mentioned mapping to each image in the dataset, the GLCM of each mentioned in experimental results.

III.˝ Experimental Results

We extracted the Haralick features form the normalized GLCM calculated for each image. We found that the values of all Haralick features differed significantly for cancerous and non-cancerous images and due to this reason, the classification of images between cancerous and non-cancerous became possible.

Later we calculated the accuracy, specificity, precision and sensitivity for SVM classifier with 3 different kernel functions. The data thus obtained is tabulated in Table I.

TABLE I Prediction efficiency parameters for 3 different kernel functions

	Sensitivity	Precision	Specificity	Accuracy
Polynomial	78.43%	100%	100%	87.78%
RBF	85.11%	100%	100%	92.22%
Linear	100%	94.34%	92.5%	96.67%

The following are the confusion matrices obtained for the individual kernel functions. Table II (A) shows the Confusion matrix for linear kernel function, Table II (B) shows the Confusion matrix for RBF kernel function and Table II (C) shows the Confusion matrix for the polynomial kernel function. We need to search for a cancerous image, keeping this in mind a positive finding is defined as an image with a non-cancerous cervix. A false negative finding is defined as a cancerous image misclassified as non-cancerous Hence it becomes imperative for the SVM Model to minimize the number of false negative findings. The Tables II(A), II (B), II (C) show that the percentage of images classified as false positive is 3.33%, 0%, 0% respectively. These values are clearly small irrespective of the kernel used. One possible method of lowering the chances of a false negative finding is by taking more number of images of the target cervix in order to detect cancer. On the similar lines, a false positive finding is said to be a non-cancerous cervix classified as cancerous. If a false positive finding occurs, the patient can be recalled for the test. In our case, the percentage of the false positive findings for the linear kernel, RBF kernel and polynomial kernel is 0%, 7.78%, 12.22% respectively. It can be observed that the percentage of false positive finding is least for the linear kernel. Fig 4 shows the efficiency of Polynomial, RBF and Linear kernel functions.

TABLE II (A)
Confusion matrix for linear kernel function

Linear Kernel function		**Predicted**	
		Negative	**Positive**
Actual	Negative	55.56%	0%
	Positive	3.33%	41.11%

TABLE II (B)
Confusion matrix for RBF kernel function

RBF Kernel function		**Predicted**	
		Negative	**Positive**
Actual	Negative	47.78%	7.78%
	Positive	0%	44.44%

TABLE II (C)
Confusion matrix for polynomial kernel function

Polynomial Kernel function		**Predicted**	
		Negative	**Positive**
Actual	Negative	43.33%	12.22%
	Positive	0%	44.44%

After achieving an accuracy of 96.67% using the linear kernel function and 13 Haralick features, we selectively removed one feature at a time and trained the SVM model. The trained model was then tested on the same testing dataset as before and we obtained the above-mentioned table of accuracy. The Table III suggests that removing entropy and energy and sum average does not affect the accuracy while using the SVM kernel.

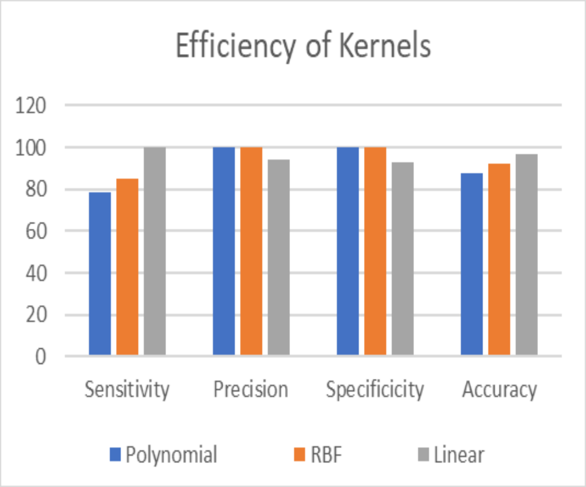

Fig. 4 Graph depicting the efficiency of kernel functions.

We chose SVM over other Machine learning algorithms because SVM is less sensitive to noise in the input while a small number of disturbances in the input can cause a huge decrease in the accuracy in case of Neural Networks. Also, problems become difficult to debug with Neural Networks because it is difficult to keep track of what is happening from a layer to another layer. Also, we did not consider K- Nearest Neighbors because they are computationally costly when used with a large dataset. We got an accuracy of 96.67% using Linear Kernel function with SVM, which is quite good.

TABLE III
ORIGINAL ACCURACY AND VARIATION OF ACCURACY AS ONE OF THE HARALICK FEATURE IS REMOVED

Original Accuracy	**96.67%**
Feature Removed	**Accuracy Obtained**
Correlation	93.55%
Energy	96.67%
Variance	91.94%
Contrast	93.55%
Homogeneity	93.55%
Sum Average	96.67%
Sum Variance	91.94%
Sum Entropy	93.55%
Entropy	96.67%
Difference Variance	95.16%
Difference Entropy	93.55%
Information Measure of Correlation I	95.16%
Information Measure of Correlation II	93.55%

IV." CONCLUSION

Cervical cancer is a disease that affects millions of women every year. In this paper, we have proposed an automated cervical cancer detection technique. The paper involves creating the GLCM and extracting the aforementioned Haralick Texture Features. Three different types of kernel functions namely Polynomial, RBF and linear kernel functions were used for classification using the SVM classifier. Our analysis showed that the Polynomial kernel function gave the accuracy of 87.78%, RBF kernel function gave the accuracy of 92.22% while linear kernel function gave the highest accuracy of 96.67%.

V." FUTURE WORK

Future work may involve image segmentation for segregating useful information from the image and implementing principal component analysis (PCA) for dimensionality reduction. Furthermore, an additional glare removal algorithm can also be applied on the image for noise removal of the images. Lange et. al. [20] propose an automated method to remove glare in reflectance imagery of the uterine cervix. This will help increase the accuracy.

ACKNOWLEDGEMENT

This research has been carried out at CSIR-Central Electronics Engineering Research Institute (CEERI), Pilani, India and Authors would like to thank Director, CSIR-CEERI for his active encouragement and support.

REFERENCES

[1]" P. D. Shahare and R. N. Giri, "Comparative analysis of artificial neural network and support vector machine classification for breast cancer detection," IRJET, vol. 02, issue 09, Dec 2015

[2]" J. Yessi, S. C. Ng, and N.A. Osman, "Intelligent screening systems for cervical cancer," The Scientific World Journal 2014, 810368, July 2018.

[3]" B. Ashok, and P. Aruna, "Comparison of feature selection methods for diagnosis of cervical cancer using SVM classifier," Journal of Engineering Research and Applications, vol. 6, issue 1, January 2016.

[4]" D. Kashyap et al., "Cervical cancer detection and classification using independent level sets and multi SVMs, "39th International Conference on Telecommunications and Signal Processing (TSP)" Vienna, 2016.

[5]" M.K. Soumya, K. Sneha, and C. Arunvinodh, "Cervical cancer detection and classification using texture analysis", Biomedical and Pharmacology Journal, 9(2):663– 71, 2016.

[6]" N. Mustafa, N. A. Mat Isa, M. Y. Mashor, and N. H. Othman, "New features of cervical cells for cervical cancer diagnostic system using neural network," IJSSST, 9(2), 2008.

[7]" P. Mohanaiah, P. Sathyanarayana and L. Gurukumar, "Image texture feature extraction using GLCM approach", International Journal of Scientific and Research Publication, vol. 3, issue 5, 2013.

[8]" S. Kaaviya, V. Saranyadevi and M. Nirmala, "PAP smear image analysis for cervical cancer detection," 2015 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, pp. 1-4, 2015.

[9]" K. Pradeep Chandran and U.V. Ratna Kumari, "Improving cervical cancer classification on MR images using texture analysis and probabilistic neural network," IJSETR, vol. 4, issue 9, 2015.

[10]" P. K. Malli and S. Nandyal, "Machine learning technique for detection of cervical cancer using k-NN and artificial neural network," IJETTCS, vol. 6, issue 4, 2017.

[11]" T. Chankong, N. Theera-Umpon, and S. Auephanwiriyakul, "Cervical cell classification using fourier transform", 13th International Conference on Biomedical Engineering, IFMBE Proceedings, vol 23, 2009.

[12]" S. Garg, S. Urooj and R. Vijay, "Detection of cervical cancer by using thresholding & watershed segmentation," 2nd International Conference on Computing for Sustainable Global Development, pp. 555-559, 2015.

[13]" T.A. Sajeena and A.S. Jereesh, "Automated cervical cancer detection through RGVF segmentation and SVM classification," International Conference on Computing and Network Communications (CoCoNet*)*, pp. 663-669, 2015.

[14]" G. Acosta-Mesa Héctor, Z. Barbara, R. Homero et al. "Cervical cancer detection using colposcopic images: a temporal approach," Sixth Mexican International Conference on Computer Science (ENC'05), pp. 158-164, 2005.

[15] Z. Meng et al. “Automatic screening of cervical cells using block image processing,” BioMedical Engineering Online, 2016.

[16] Zhang J. and Liu Y., “Cervical cancer detection using SVM based feature screening”, In: Barillot C., Haynor D.R., Hellier P. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004, Lecture Notes in Computer Science, vol 3217, 2004.

[17] P.P. Ohanian and R.C. Dubes “Performance evaluation for four classes of textural features”, Pattern Recognition, 25(8), pp. 819–833, 1992.

[18] C. Cortes and V. Vapnik “Support-vector networks” Machine Learning, Kluwer Academic Publishers Boston, 20(3), pp. 273–297, 1995.

[19] R. Haralick, K. Shanmugam, and D. Itshak. “Textural features for image classification”, IEEE Trans Syst Man Cybern 3, pp. 610-621, 1973.

[20] L. Holger, “Automatic glare removal in reflectance imagery of the uterine cervix,” Proceedings of SPIE - The International Society for Optical Engineering, 2005.