Volume 6, Issue 11, November 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Robust Diagnosing Technique for Cervical Cancer Using Random Forest Classifier

B. Ashok, Dr. S. Anu. H. Nair, N. Puviarasan, Dr. P. Aruna
Department of Computer Science&Engineering, Annamalai University, Annamalainagar, Tamilnadu, India

Abstract— The main aim of this research work is to find an easy and accurate method to diagnose the cervical cancer which is the one of the deadliest cancers among women. Image processing techniques are applied on the cervical cell image which is obtained through pap smear test. Further, to optimize the features, feature selection process using genetic algorithm is performed. Random forest classifier is utilized to differentiate the normal and abnormal cancer cells. The result reveals that the best quality of diagnosis of cervical cancer. This work can give benefit to pathologists to diagnose the disease.

Keywords— GLCM, LBP, Active contour Model, Genetic Algorithm, Feature selection, Random Forest, Cervical Cancer

I. INTRODUCTION

Cervical cancer is one of the silent killing disease among women society. Cervical cancer can be preventable. Major cause for cervical cancer is the infection of Humon Papilloma Virus (HPV) which is a sexually transmitted virus. There are more than 100 types of HPV viruses exist but HPV 16 and HPV 18 are highly related to the cause of cervical cancer. Early detection of the disease can preserve the life of the patient. For the detection of abnormalities of cervix require screening methods or biopsy test. Biopsy need surgery which is expensive, painful and time consuming process. In 1940s, Papanicolaou discovered that vaginal smears indicate the growth of the cancer cells. This test is known as pap smear test which is a easy and noninvasive test. In the western countries, frequent screening tests are conducted for the women so HPV infected women rate is also less. In the developing countries, HPV affected women rate is nearly 85% due to lack of screening test.

In this paper, textural features are obtained from pap smear images to diagnose the cervical cancer. Textural features are extracted using GLCM, LBP and Tamura methods. Feature selection is carried out by genetic algorithm. Random forest classifier is used to decide whether it is normal or abnormal cancel cell.

The paper is organized as follows. Section 2 describes literature review. Methodologies are discussed in section 3. Result and discussion placed in section 4. Conclusion is in section 5.

II. LITERATURE REVIEW

Papanicolaou G.N. et al (1943) discovered that cell abnormalities can be identified from vaginal smears. Sayeda, Ahmed M., et al (2016) made classification of breast tumors using magnetic resonance images. They extracted features from histograms and applied kNN classifier for tumor classification. Lingzheng Dai et al (2015) utilized active contour model for natural image segmentation. Lutful Mabood et al (2016) performed image segmentation based on active contour model. This model was capable to selectively segment and capture objects with non uniform features. Mellisa Pratiwi et al (2015) proposed the classification method for breast images based on GLCM textural features. They compared two classifiers such as Back-Propagation Neural Network (BPNN) and Radial Basis Function Neural Network (RBFNN).

Michael Kass et al (1988) elaborately explained about active contour models for segmenting image objects such as lines and edges. Karthigai Lakshmi, G. et al (2016) conducted experiments for classification of cervical cell images using SVM classifier. Bichen Zheng et al (2014) investigated about breast cancer diagnosis based on feature extraction by combining k-means and support vector machines. John Arevalo et al (2015) explained the method for detection of cancer from histopathology images. Nongyao Nai-arun et al (2015) suggested that best classification result achieved from random forest classifier. Amira Sayed A. Aziz et al (2013) applied genetic algorithm for feature selection and obtained best feasible feature subset. Manju B et al (2016) conducted experiments for diagnosing prostate related diseases using CT images. They applied genetic algorithm to optimize the features. Monjoy Saha et al (2016) discussed about the diagnosis of breast cancer using cytological images. Siti Noraini Sulaiman et al (2015) presented feature extraction methods and classification of cervical cancer using neural networks. They applied the Adaptive Fuzzy-k-Means (AFKM) clustering algorithm to replace the Moving k-Means (MKM) to segment pap smear images into the nucleus, cytoplasm and background regions. Bhuvaneswari C et al (2013) performed classification of lung diseases using CT images and applied genetic algorithm for feature selection. Intan Aidha Yusoff et al (2010) performed cervical cell classification using multilayered perceptrons.

III. METHODOLOGY

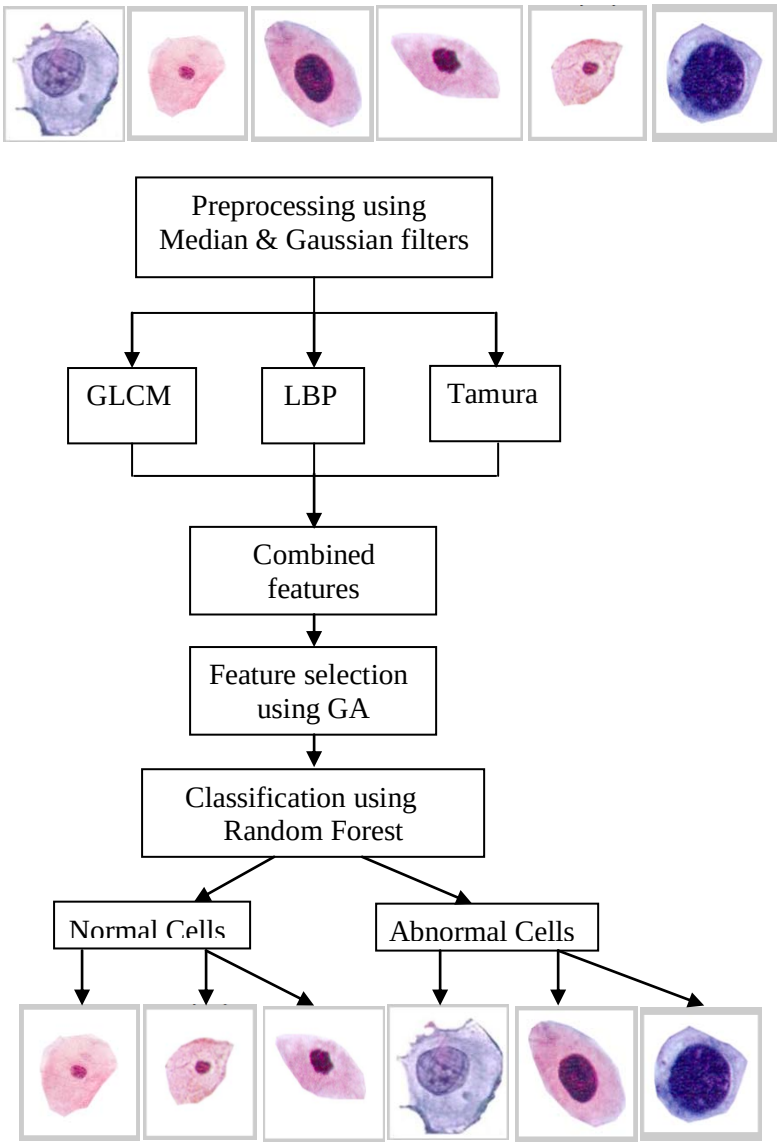

Fig. 1 Block diagram of the proposed work

Fig. 1 shows the block diagram of this work. Preprocessing of input images is the first stage. GLCM, LBP and Tamura features are extracted. All these features are combined. Feature selection and classification are performed.

A. Dataset

To implement the method of this research work, the dataset is obtained from Rajah Muthiah Medical College & Hospital, Annamalai University, Annamalainagar. The dataset contains 479 cervical cell images.

B. Pre-processing

Image preprocessing is concerned with enhancement of visual appearance of an image. Preprocessing consists of the image enhancement, noise removal, smoothening and image resampling. In this work, median filter is applied for removal of noise and then Gaussian filter is applied for smoothening the cervical cell image.

The median filter is mentioned as

$$Y(t) = median((x\left(t - \frac{T}{2}\right), x(t - T_1 + 1), \dots x(t), \dots x\left(t + \frac{T}{2}\right)) \quad (1)$$

where Y is median filter, t is size of the window of the median filter, x represent the image, T is pixel location and T_1 is pixel at first row and first column (r_1, c_1).

Gaussian filter is the most common filter used to smoothen the image in preprocessing stage of various medical applications by researchers. In 2-D, an isotropic Gaussian is denoted as follows:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{\frac{-x^2+y^2}{2\sigma^2}} \quad (2)$$

where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis and σ is the standard deviation of the Gaussian distribution.

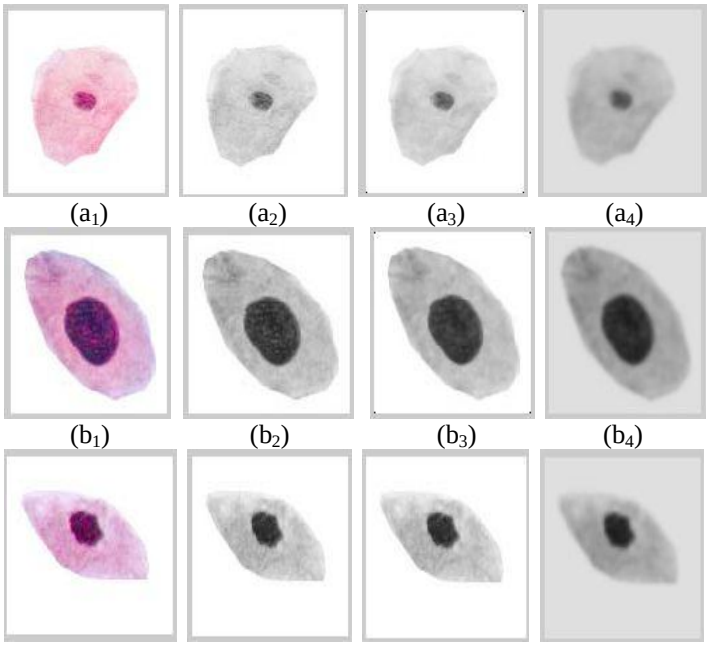

Fig. 2 Image preprocessing stages – sample images

Fig.2 shows the various stages in image preprocessing. (a_1, b_1 and c_1) are the original images, (a_2, b_2 and c_2) are the gray scale images, (a_3, b_3 and c_3) are the median filtered images and (a_4, b_4 and c_4) are the smoothened images using Gaussian filter.

C. Feature extraction

Feature extraction is the process of obtaining information from the image object for classification. Texture features in image are referred by the principle of human visual systems of natural things and the existence of basic primitives. In this work, texture features are obtained using Gray level co-occurrence matrix (GLCM), Local binary pattern (LBP) and Tamura methods.

1) GLCM

The GLCM is a matrix of how different combinations of pixel gray levels occur in an image. The number of rows and columns in the GLCM is equal to number of gray level in the corresponding image. Spatial relationship between the pair of pixels in the image and the distance are considered for GLCM. For an image $I(x, y)$ of size $Mx\ M$ the co-occurrence matrix P can be defined as

$$P(i,j) = \sum_{x=1}^{M} \sum_{y=1}^{M} \begin{cases} 1, if\ I(x,y) = i\ and\ I(x + \Delta_{x,y} + \Delta_y) = j \\ 0, otherwise \end{cases} \quad (3)$$

where $\Delta_x \Delta_y$ are the distance between pixel of interest and its neighbor pixel.

14 textural features are derived from the input images such as homogeneity, contrast, correlation, entropy, variance, inverse difference moment, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure I, information measure II and maximal correlation coefficient (Haralick, R.M. et al 1973). Four different angles such as 0^0, 45^0, 90^0 and 135^0 are implemented to get corresponding four co-occurrence matrices. Totally, 56 GLCM features are obtained by 14 textural features for four different angles.

2) LBP features

LBP is a texture feature which calculates the pixel intensities in a local neighborhood (Xiangping Sun et al 2011).

The binary pattern of LBP value for any location $(x, y)$of an image could be written as

$$LBP_{P,R} = \sum_{p=0}^{P-1} s\left(g_p - g_C\right) 2^P \quad (4)$$

where g_p represents intensity value of any neighbouring pixel centered around g_c indicating the pixel value of the center pixel at (x, y). LBP is calculated by summing the threshold differences weighted by powers of two.

In our work, the gray scale image is subdivided into non-overlapping 3 x 3 partitions as 9 sub images. Each sub image contains one LBP value. Totally, 9 LBP values are obtained from each image.

3) Tamura features

Hideyuki Tamura et al., (1978) proposed texture properties based on psychological studies of human visual perception which are very useful for parameterization of appearance of object and its subsequent recognition. These properties consist of six statistical features, which include coarseness, contrast, directionality, regularity, line-likeness, and roughness. Tamura features can be calculated for a whole image or sub images. Extracting tamura features, for a very small image may invalidate the statistical features. In this work, tamura features are calculated for the input image which is divided into non overlapping 3 x 3 sub images. From each sub image, six tamura features are extracted. Totally, 54 Tamura features are obtained for each image.

4) Proposed combined Texture features

The GLCM features $G = \{g_1, g_2 \dots g_n\}$, LBP features $L = \{l_1, l_2 \dots l_m\}$ and Tamura features $T = \{t_1, t_2 \dots t_k\}$ are combined to form the proposed combined texture features (CT) which is defined as

$$CT = (G + L + T) \quad (5)$$

In this feature extraction phase, 56 GLCM features, 9 LBP features and 54 Tamura features are obtained. Hence, 119 weighted textural features are totally obtained.

D. Feature selection

In this work, Genetic Algorithm (GA) is applied for feature selection. The genetic algorithm is based on Darwinian's theory of survival of the fittest. Genetic algorithm is wrapper based and effective feature selection technique. It is an iterative procedure and each iteration is known as generation. Mating is the main operation in which two solutions are mated to produce a new solution. Mating is performed by two operators, called crossover and mutation. Genetic algorithm functionality is explained in Algorithm 1

Algorithm 1 Genetic Algorithm

Input : Full feature set
Output: Optimized feature subset
Step 1: [start] Start with initial population of N chromosomes randomly.
Step 2: [Fitness] A fitness function $F(x_i) = \frac{f(x_i)}{\sum_{i=1}^{N_{ind}} f(x_i)}$ is applied for evaluation.
Step 3: [New population] Next generation of population is produced.
Step 4: [Selection] In this process, the best offsprings are selected according to their fitness from the current generation.
Step 5: [crossover] Crossover operator is applied by swapping the genes at one or more points to produce a new offpring.
Step 6: [Mutation] Mutation operator is also applied to produce a better offspring.
Step 7: [Accepting] The new offspring is placed in the mating pool.
Step 8: [Replace] Better offsprings are replaced by the old ones.
Step 9: [Evaluate] Evaluate population.
Step10: [Check] If the termination criterion is reached, then the best solution is obtained.
Step 11: [Else] Go to step 3.

E. Random Forest

The random decision forests were first created by Tin Kam Ho then it was developed by Leo Breiman and Adele Cutler. Random Forest is an ensemble classifier and gives of higher rate of accuracy. Random Forest is created using more number of multiple binary decision trees. The trees are growing with the training data with the target of avoiding the over-fitting problem of individual decision trees. Over-fitting means tuning the classifier very close to the training data which prevent the accuracy of the classifier. Decision trees are combined individual learners.

Random Forest algorithm is based on CART (Classification and Regression Tree) which is one type of decision trees. These trees are greedy and split nodes in recursive partitioning basis from top down manner. Two third of the cases of data are considered for training and one third of data for misclassification error. Predictor variable is randomly selected and used to spilt the nodes, for each tree calculate misclassification error based on out of bag (OOB) error rate using the one third of data and aggregating the error from all the trees gives the overall OOB error rate for classification. Each tree gives a classification, for a binary variable 0 or 1. The forest counts the classification results and selects the bigger value. Random forest algorithm is the statistical and machine learning algorithm. It uses multiple learning algorithms for better performance. (Pradeep Kandhasamy, J., et al 2015). This algorithm has two parts such as tree bagging and bagging to random forest.

Each tree in the forest is trained as follows:

1. From the original data randomly sample N cases with replacement where N is the number of cases in the training set. Growth of the tree is ensured by the sample cases.
2. Even though the attributes are selected randomly, the number of attributes is kept constant while the trees are growing and the best split method is used to split the node.
3. Each tree is grown without pruning to the maximum level. There is no pruning. Pruning means that to avoid over-fitting and to achieve accuracy and stability.

The forest error rate depends on the two things

1. The correlation between any two trees in the forest. Increasing the correlation improves the forest error rate.
2. The stability of each and every tree in the forest. Good classifier has a low error rate from the tree. Increasing the stability of the individual trees decreases the forest error rate.

The node splitting criteria is given using Gini impurity. Let N be the root node and it got split into two child nodes C_{left} and C_{right}. If C_{left} and C_{right} are too small and are not able to further split, make it as leaf nodes L_{left} and L_{right} else make it as left root node N_{left} and N_{right} right root node. Repeat the procedure until the child becomes leaf node.

GINI criterion for splitting nodes is given as

$$Gini = N_L \sum_{k=1}^{n} p_{kL}(1 - p_{kL}) + N_R \sum_{k=1}^{m} p_{kR}(1 - p_{kR}) \quad (6)$$

where p_{kL} = proportion of class k in left node

p_{kR} = proportion of class k in right node

IV. RESULTS AND DISCUSSION

Table 1 shows the selected number of features using genetic algorithm. In GLCM method, out of 56 features, 31 feature are selected. LBP method contains only 9 features. Therefore, selection is not performed for this method. In tamura method, out of 54 features, 34 features are chosen using GA method. For the combined textural features, out of 119 features 63 features are chosen by the GA feature selection method.

Table 1 Feature selection using Genetic Algorithm

Sl.no	Feature extraction methods	Total features	Feature selection using GA
1	GLCM	56	31
2	Tamura	54	34
3	Combined Textural features (including LBP features)	119	63

The performance of the random Forest classifier is measured in terms of accuracy, sensitivity and specificity.

1) Accuracy: Accuracy is obtained by correctly classified images divided by the classified images.

$$\text{Accuracy} = \frac{TP + TN}{(TN + TP + FP + FN)} \quad (7)$$

where TP is True positive, TN is True negative, FP is False positive and FN is False negative

Table 2 Accuracy of Feature extraction methods

Sl. no	Feature extraction method	Accuracy(%)
1	GLCM	62.35
2	LBP	59.42
3	Tamura	60.36
4	Combined Texture	81.71

Table 2 shows the accuracy of various feature extraction methods using random forest classifier. The maximum level of accuracy is obtained by the combined texture features.

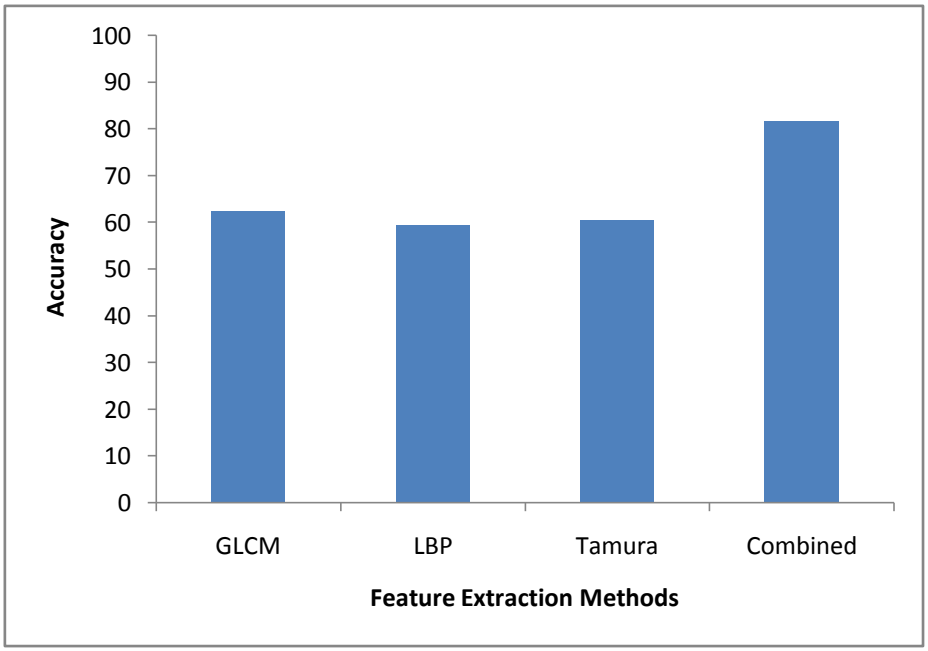

Fig. 3 Accuracy of Texture Features using Random Forest classifier

Fig. 3 represents the accuracy of various feature extraction methods such as GLCM, LBP, Tamura and combined features of all these methods. The feature selection is performed individually for the GLCM, Tamura and combined texture features and not performed for LBP features. The accuracy of GLCM method get 62.35%, LBP method 59.42%, Tamura method 60.36% and combined texture feature method 81.71%. So, the better accuracy is achieved from the combination of texture method than the other individual methods.

2) Sensitivity: Sensitivity is obtained as correctly classified true positive rate divided by true positive and false negative samples.

$$Sensitivity = \frac{TP}{TP+FN} \quad (8)$$

Table 3 Sensitivity of Feature extraction methods

Sl. no	Feature extraction method	Sensitivity(%)
1	GLCM	61.39
2	LBP	58.94
3	Tamura	60.79
4	Combined Texture	80.48

Table 3 shows the sensitivity of various feature extraction methods using random forest classifier. The maximum level of sensitivity is obtained by the combined texture features as 80.48%.

3) Specificity: Specificity is calculated as correctly classified true negative rate divided by the true negative and false positive samples. Results that are true negatives are treated as errors.

$$Specificity = \frac{TN}{TN+FP} \quad (9)$$

Table 4 Specificity of Feature extraction methods

Sl. no	Feature extraction method	Specificity(%)
1	GLCM	61.77
2	LBP	57.28
3	Tamura	59.62
4	Combined Texture	80.15

Table 4 shows the specificity of various feature extraction methods using random forest classifier. The maximum level of specificity is obtained by the combined texture features as 80.15%.

V. CONCLUSION

In this paper, image processing techniques such as image preprocessing and feature extraction are analysed. Machine language techniques such as feature selection and classification of cervical cell are also analysed. In feature selection phase, Genetic Algorithm is applied to get the optimized feature subset. Combination of textural features with Random forest classifier gives better accuracy results than individual results of GLCM, Tamura and LBP methods. The accuracy, sensitivity and specificity for different feature extraction method using Random forest classifier are compared and found that combined texture features gives better result than other methods with the value of 81.71%,

REFERENCES

[1] Amira Sayed A. Aziz, Ahmad Taher Azar, Mostafa A. Salama, Aboul Ella Hassanien and Sanaa El-Ola Hanafy (2013), Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors Generation, *Computer Science and Information Systems*, pp. 769–774

[2] Bhuvaneswari C., Aruna P. and Loganathan D. (2013), Classification of the Lung Diseases from CT Scans by Advanced Segmentation Techniques using Genetic Algorithm, *International Journal of Computer Applications*, Vol. 77 – No. 16 , pp 21-27.

[3] Bichen Zheng, Sang Won Yoon and Sarah S. Lam (2014), Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, *Expert Systems with Applications, (Elsevier)*, Vol. 41, pp 1476–1482.

[4] Haralick, R.M., Shanmugam,K. and Dinstein, I. (1973), Textural Features for Image Classification , *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 3, no. 6, pp. 610–621.

[5] Hideyuki Tamura, Shunji Mori and Takashi Yamawaki (1978), Textural Features Corresponding to Visual Perception, *IEEE Transactions on Systems. Man and Cybernetics*, Vol. SMC-8, No. 6, pp 460 – 473.

[6] Intan Aidha Yusoff, Nor Ashidi Mat Isa, Nor Hayati Othman, Siti Noraini Sulaiman and Yessi Jusman (2010), Performance of neural network architectures: Cascaded MLP versus extreme learning machine on cervical cell image classification, *Signal Processing and their Applications (ISSPA 2010)*, pp 308-311.

[7] John Arevalo, Angel Cruz-Roa, Viviana Arias, Eduardo Romero and Fabio A. Gonzalez (2015), An unsupervised feature learning framework for basal cell carcinomaimage analysis, *Artificial Intelligence in Medicine,(Elsevier)*,Vol.64, pp 131 – 145.

[8] Karthigai Lakshmi, G. and Krishnaveni, K. (2016), Feature Extraction and Feature Set Selection for Cervical Cancer Diagnosis, *Indian Journal of Science and Technology*, Vol. 9(19).

[9] Lingzheng Dai, Jundi Ding and Jian Yang (2015), Inhomogeneity-embedded active contour for natural image segmentation, *Pattern Recognition, (Elsevier)*, Vol. 48, Issue 8, pp 2513-2529.

[10] Lutful Mabood, Haider Ali, Noor Badshah, Ke Chen and Gulzar Ali Khan (2016), Active contours textural and inhomogeneous object extraction, *Pattern Recognition, (Elsevier)*, In Press, Accepted Manuscript, Available online 3 February 2016.

[11] Manju B, Meenakshya,K. and Gopikakumari,R. (2015), Prostate Disease Diagnosis from CT images using GA optimized SMRT based Texture Features, *Procedia Computer Science,(Elsevier)*, Vol. 46 pp 1692 – 1699.

[12] Mellisa Pratiwi, Alexander, Jeklin Harefa and Sakka Nanda (2015), Mammograms Classification using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network, *Procedia Computer Science,(Elsevier)*, Vol. 59 pp 83 – 91.

[13] Michael Kass, Andrew Witkin and Demetri Terzopoulous (1988), Snakes: Active Contour models, *International Journal of Computer vision*, pp 321-331.

[14] Monjoy Saha, Rashmi Mukherjee and Chandan Chakraborty (2016), Computer-aided diagnosis of breast cancer using cytological images: A systematic review, *Tissue and Cell, (Elsevier)*, Volume 48, Issue 5, pp 461-474.

[15] Nongyao Nai-arun and Rungruttikarn Moungmai (2015), Comparison of Classifiers for the Risk of Diabetes Prediction, *Procedia Computer Science,(Elsevier)*, Vol. 69, pp 132 – 142.

[16] Papanicolaou G.N., Traut H.F., Stanton M. and Friedberg A.(1943), Diagnosis of Uterine Cancer by the Vaginal Smear, *Oxford University Press*, New York.

[17] Pradeep Kandhasamy, J., and Balamurali, S. (2015), Performance Analysis of Classifier Models to Predict Diabetes Mellitus, *Procedia Computer Science, (Elsevier)*, Vol.47, pp 45 – 51.

[18] Sayeda, Ahmed M., Eman Zaghloulb and Nassef, Tamer M. (2016), Automatic Classification of Breast Tumors Using Features Extracted from Magnetic Resonance Images, *Procedia Computer Science, (Elsevier)* pp 392 – 398.

[19] Siti Noraini Sulaiman, Nor Ashidi Mat-Isa, Nor Hayati Othman and Fadzil Ahmad (2015), Improvement of Features Extraction Process and Classification of Cervical Cancer for the NeuralPap System, *Procedia Computer Science, (Elsevier)*, Vol.60, pp 750 – 759.

[20] Xiangping Sun, Jin Wanga, Ronghua Chen, Lingxue Kong and Mary F.H. She (2011), Directional Gaussian Filter-based LBP Descriptor For Textural Image Classification, Advanced in Control Engineering and Information Science, *Procedia Engineering, (Elsevier)*, Vol. 15, pp 1771 – 1779.

