RESEARCH ARTICLE OPEN ACCESS

Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier

B. Ashok*, Dr. P. Aruna**
*(Department of Computer Science & engineering, Annamalai University, Annamalainagar, India)
** (Department of Computer Science & engineering, Annamalai University, Annamalainagar, India)

ABSTRACT

Even though a great attention has been given on the cervical cancer diagnosis, it is a tuff task to observe the pap smear slide through microscope. Image Processing and Machine learning techniques helps the pathologist to take proper decision. In this paper, we presented the diagnosis method using cervical cell image which is obtained by Pap smear test. Image segmentation performed by multi-thresholding method and texture and shape features are extracted related to cervical cancer. Feature selection is achieved using Mutual Information(MI), Sequential Forward Search (SFS), Sequential Floating Forward Search (SFFS) and Random Subset Feature Selection(RSFS) methods.

Keywords - Segmentation, Feature Extraction, Feature Selection, Multi-thresholding, Classification

I.INTRODUCTION

Cervical cancer is the fourth common cancer in women society. In the low-income countries, the acceptability of increased cervical cancer risk and cause of cancer death projected. Cervical screening programs has minimized the death rate in the developed countries. Human Papilloma Virus (HPV) is found as the major cause for the cervical cancer. Although cervical cancer is one of the deadliest disease, it can be cured if detected in the earlier stage. Pap smear test is the non-invasive screening and simple test to find out cervical cancer. Image acquisition is obtained by applying the smear which is taken by the pap smear test on the microscope slide. From microscope attached digital camera capture the cell image. Image processing is the major area that includes image enhancement, segmentation and feature extraction. Feature selection and classification are the machine learning techniques.

In this paper we have applied image processing, data mining and machine learning techniques to diagnose the cervical cancer.

II.RELATED WORK

Feature selection is an important process. The feature selection methods can be divided into filter, wrapper and embedded methods [1,2]. Filter methods are computationally very fast, easy to interpret and are scalable for high- dimensional datasets like micro array data. However, most of them are univariate. Wrapper and embedded methods which always use machine learning algorithms to compute the importance scores of the feature or feature subset, RF-ACC and RF-Gini based on Random Forest(RF) [3] always achieve better performances than filter methods on high-dimensional datasets. Premature time diagnosis of the disease could improve patients' treatment effect. One of the most important and indispensable tasks in any pattern recognition system is to overcome the curse of dimensionality problem, which forms a motivation for using a suitable feature selection method [4]. Random based feature selection[5] using benchmark datasets also discussed and a frame work of algorithmic randomness based feature selection has been proposed to measure the feature importance score using a p-value that derives from the combination of algorithmic randomness test and machine learning methods.

Rough Set Theory (RST) and Particle Swarm Optimization (PSO) methods applied for medical diagnosis and achieved accuracy to certain extent [7]. Canonical Correlation Analysis (CCA) for uncertain data streams and an uncertain canonical correlation analysis (UCCA) method discussed to reduce the dimensionality of the features [8]. Extracting shape features, textural features and intensity features [9] was achieved by watershed transformation and k-means spectral clustering supervised svm classifiers applied for feature selection. Multi-Filtration Feature Selection (MFFS) [10], a approach which was applied to extract features, feature subset selection, feature ranking and classification.

A cosine similarity measure support vector machine CSMSVM [11] for feature selection and classification presented. Feature selection is performed by adding weights to features based on margin verses cosine distance ratio. An unsupervised feature learning (UFL) framework for basal cell carcinoma image analysis[12] had been performed on histopathology images.

The K-SVM had been explained [13]. Missing of important data and gathering unnecessary information at the time of feature extraction should be avoided. Multi resolution local binary pattern (MRLBP) [14] variants based texture feature extraction techniques had been presented. Image segmentation performed using discrete wavelet transform (DWT) further Daubechies wavelet(db2) used as decomposition filter. Deep Sparse Support Vector Machine (DSSVM) based feature selection method had been discussed [15] and further extraction of color and texture features was performed using super pixels. The importance of local texture(LTP) and local shape (MultiHPOG) features discussed [16] and the further improvement in performance achieved by adding the global shape feature (Gabor wavelets). Feature selection based on information has the drawback such as the interaction between the features and classifiers. This may lead to the selection irrelevant features. To overcome this problem, Mutual Information based techniques used [17]. Even though new feature extraction, feature selection and classification methods have been introduced, there will be a requirement exists for newer or better techniques. In our previous works [18][19] various segmentation methods were analyzed.

III.MATERIALS AND METHODS

The Block diagram of the proposed method

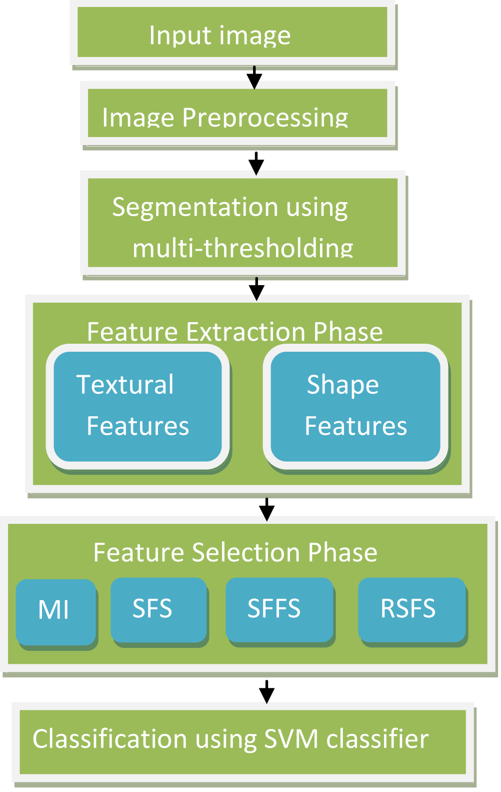

Figure 1: Block Diagram of the proposed method for diagnosis of cervical cancer

Image preprocessing includes noise removal and image enhancement. Image resize also performed. Cell image converted from RGB image to Gray scale image.

3.1 Image segmentation

Segmentation means split the input image into desired regions so as to get the features. There are lot of methods to segment an image such as edge based methods, region based, watershed method and thresholding methods. In this work, we used multi thresholding method to segment the input image. Nucleus and Cytoplasm of the cervical cell image are segmented.

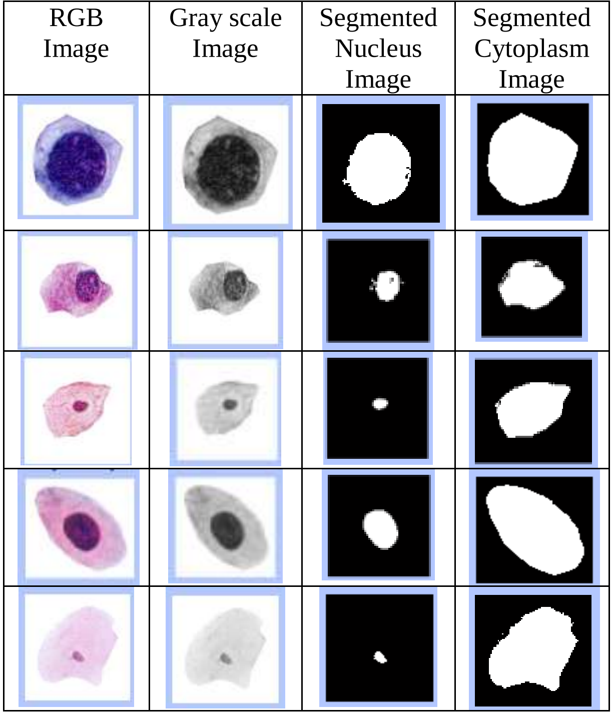

Figure 2: Sample pap smear images & their corresponding segmented images

3.2 Feature Extraction

3.2.1 Texture features

Texture features are extracted using Gray Level Co-occurrence Matrix (GLCM). The list of features are as follows homogeneity, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measure I, information measure II and maximal correlation coefficient. Totally 14 texture features are extracted.

3.2.2 Shape features

The following list of features are the shape features which are extracted from nucleus and cytoplasm of the cervical cell. Besides the shape features, derived features are also included. Nucleus related features are Area, Centroid, Eccentricity, Equiv Diameter, Major Axis Length, Minor Axis Length, Perimeter, roundness, position, brightness and nucleus cytoplasm ratio and cytoplasm related features are Area, Centroid, Eccentricity, Equiv Diameter, Major Axis Length, Minor Axis Length, Perimeter, brightness, roundness. Totally 30 shape features are extracted. In this proposed work, by combining texture and shape features we extract totally 44 features.

IV.Feature Selection

Feature selection is one of the optimization process to select a minimum number of effective features. Feature selection is required to reduce curse of dimensionality due to irrelevant or overfitting, cost of feature measurement and computational burden. In order to select a feature subset, a search method and an objective function is required. Small sample size and what objective function to use are some of the potential difficulties in feature selection process. Number of features is directly proportional to the feature subset that is as in the case of an increase in features there may be an increase in feature subset also.

There is a large number of search methods, which can be grouped in three categories. Exponential algorithms, Sequential algorithms and Randomized algorithms. Exponential algorithms evaluate a number of subsets that grows exponentially with the dimensionality of the search space. Exhaustive Search, Branch and Bound, Beam Search comes under exponential algorithms. Sequential algorithms add or remove features sequentially, but have a tendency to become trapped in local minima. Representative examples of sequential search include Sequential Forward Selection, Sequential Backward Selection, Sequential Floating Selection and Bidirectional Search. Randomized algorithms incorporating randomness into their search procedure to escape local minima. Examples are Random Generation plus Sequential Selection, Simulated Annealing and Genetic Algorithms.

Objective function is divided into three types. One is filter method, second one is wrapper method and another is embedded method. Filter methods are defined as independent evaluation of the classification algorithm and the objective function evaluates feature subsets by their information content, typically interclass distance, mutual information or information-theoretic measures. In wrapper method, the objective function is a pattern classifier, which evaluates feature subsets by their predictive accuracy (recognition rate on test data) by statistical re-sampling or cross validation. Embedded methods are specific methods. In this paper filter based Mutual Information (MI) and wrapper based methods such as Sequential Forward Selection, Sequential Floating Forward selection(SFFS) and Random Subset Feature Selection methods are used to select the features.

3.3 Mutual Information (MI)

Basically Mutual information is the measurement of similarity. The mutual information of two random variables is a measure of the variable's mutual dependence that is correlation between two variables. Increase in mutual information which is often equivalent to minimizing conditional entropy. MI comes under filter method. Generally filter methods provide ranking to the features. Choosing the selection point based on ranking is performed through cross-validation. In feature selection, MI is used to characterize the relevance and redundancy between features.

The mutual information between two discreet random variables X, Y jointly distributed according to p(x, y) is given by

$$I(X;\,Y) \;=\; \sum_{x,y} p(x,y)\ \log\frac{p(x,y)}{p(x)p(y)} \qquad (1)$$

$$=\; H(X) - H(X|Y) \qquad (2)$$

$$=\; H(Y) - H(Y|X) \qquad (3)$$

$$=\; H(X) + H(Y) - H(X,\,Y) \qquad (4)$$

Where H(X) and H(Y) are the individual entropy and H(X|Y) and H(Y|X) are conditional entropy. H(X,Y) Joint entropy. p(x,y) joint distribution and p(x) probability distribution

In this work, all 44 features are ranked by the MI algorithm. The first top ranked 10 features are selected from the ranking list.

3.4 Sequential Forward Selection (SFS)

Sequential forward selection is one of the deterministic single-solution method. Sequential forward selection is the simplest and fastest greedy search algorithm start with a single feature and iteratively add features until the termination criterion is met. The following algorithm explain the concept of SFS

Algorithm

Input: the set of all features, F = $f_1,f_2,...,f_n$

Initialization: X_0 ="null set", k = 0

Addition: add significant feature
X_j = argmax $[J(X_k + X_j)]$, $X_j \in F - X_k$
where X_j is the new feature to be added
$X_{k+1} = X_k + X^+k = k + 1$
Go to Addition
Termination: stop when 'k' equals the number of desired features
Output : a subset of features, X_k

In this work, 8 features are selected by the SFS method out of 44 features. The selected feature subset is given in table 2.

Table 2: Selected Feature Numbers by SFS

Serial number	1	2	3	4	5	6	7	8
Selected feature numbers	3	5	4	6	29	18	44	20

3.5 Sequential Floating Forward Search (SFFS)

The sequential floating forward search method is one of the most effective feature selection technique. This algorithm starts with a null feature set. Then the best feature in the feature set that satisfies some criterion function is added with the current feature set. This is one process of the sequential forward selection. Further any improvement of the criterion is searched if some feature is excluded. By this way the worst feature according to the criterion is eliminated from the feature set. This is one process of sequential backward selection. In this method the algorithm continues by increasing or decreasing the number of features until the desired feature subset is obtained. The word floating denotes the increase or decrease in dimensionality of features in the selected feature subset.

Algorithm

Step 1: Inclusion. Use the basic sequential feature selection method to select the most significant feature with respect to feature set and include it in the feature subset. Stop if desired features have been selected, otherwise go to step 2.

Step 2: Conditional exclusion. Find the least significant feature *k* in feature subset. If it is the feature just added, then keep it and return to step 1. Otherwise, exclude the feature *k*. Note that feature subset is now better than it was before step 1. Continue to step 3.

Step 3: Continuation of conditional exclusion. Again find the least significant feature in feature subset. If its exclusion will leave feature subset with at least 2 features, then remove it and repeat step 3. Whenever this condition finish, return to step 1.only 6 features are selected by the sequential floating forward selection method and it is given in table 3.

Table 3 Selected Feature Numbers by SFFS

Serial number	1	2	3	4	5	6
Selected feature numbers	5	4	6	29	3	9

3.6 Random subset feature selection (RSFS)

Features are selected by classifying the data with a classifier while applying randomly chosen subsets. According to the classification performance the relevance of each feature is computed. Based on the average usefulness each feature is evaluated in random subset feature selection. 12 features are selected by the random subset feature selection method and it is given in table 4.

Table 4 Selected Feature Numbers by RSFS

1	2	3	4	5	6	7	8	9	10	11	12
3	5	28	20	9	44	12	33	14	24	34	18

V.Classification

Support Vector Machine (SVM) is supervised learning method and one of the kernel- based techniques in machine learning algorithms. The SVM is developed mainly by Vapnik at Bell laboratories. Basically SVM constructs a high-dimensional space and then separates according to the training data. In figure 3, the SVM classification process is illustrated. The rounded objects are support vectors.

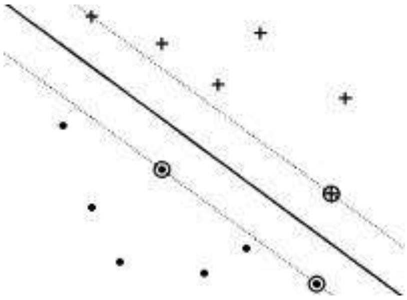

Figure 3, SVM classifier

VI.Results and Discussion

In this work, 150 images of pap smear test are collected from Rajah Muthiah Medical College, Annamalainagar, out of 150 images, 100 images are used for training the SVM classifier and 50 images are used for testing.

The combination of texture and shape features are extracted from each images. It has 44 features. The optimal features are selected using feature selection methods like Mutual Information, Sequential forward selection, Sequential floating forward selection and Random subset feature selection. The cervical cancer affected images are found using SVM classifier. Different feature selection methods are compared to find the best feature selection method suited for the diagnosis of cervical cancer.

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

$$\text{Accuracy} = \frac{\mathrm{TN}+\mathrm{TP}}{TN+TP+FN+FP} \quad (6)$$

Where TN is true negative,TP is true positive, FP is false positive and FN is false negative.

Table 5 Comparison of Accuracy of feature selection methods

Slno	Method	Accuracy %
1	MI	92 %
2	SFS	93 %
3	SFFS	98.5%
4	RSFS	94%

2) Sensitivity: Sensitivity is obtained as correctly classified true positive rate divided by true positive and false negative samples. Inconclusive results that are true positives are treated as errors for calculation.

$$\text{Sensitivity} = \frac{\mathrm{TP}}{TP+FN} \quad (7)$$

Table 6 Comparison of Sensitivity of feature selection methods

Slno	Method	Sensitivity %
1	MI	89 %
2	SFS	94.5 %
3	SFFS	98%
4	RSFS	91%

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.

$$\text{Specificity} = \frac{\mathrm{TN}}{TN+FP} \quad (8)$$

Table 7 Comparison of Specificity of feature selection methods

Sl no	Method	Specificity %
1	MI	91 %
2	SFS	92 %
3	SFFS	97.5%
4	RSFS	93 %

VII.CONCLUSION

From the table 1 to table 4, it is found that Sequential floating forward selection method outperforms the other method. That is the feature selected from Sequential floating forward selection method gives the most optimal features to diagnosis the cervical cancer. Because accuracy 98.5%, sensitivity 98% and specificity 97.5% obtained from sequential floating forward selection method which is higher than other methods.

REFERENCES

[1] Isabelle Guyon, AndreElisseeff, An introduction to variable and feature selection, *J.Mach. Learn. Res. 3 (2003) 1157–1182.*

[2] YvanSaeys, InakiInza, PedroLarranaga, A review of feature selection techniques in bioinformatics, *Bioinformatics 23(19) (2007) 2507–2517.*

[3] L. Breiman, Random forests, *Mach. Learn. 45 (1) (2001) 5–32.*

[4] N. Rami, N. Khushaba, A. Al-Ani, A. Al-Jumaily, Feature subset selection using differential evolution and a statistical repair mechanism, in: *Expert Systems with Applications, Elsevier, 2011, pp. 11515–11526.*

[5] Huazhen Wang, BingLv, FanYang , KaiZheng, XuanLi, XueqinHu, Algorithmic randomness based feature selection for traditional Chinese chronic gastritis diagnosis, *Neuro computing, Elsevier 140 (2014) pp. 252–264.*

[6] Maria Martinsa, Lino Costab, Anselmo Frizerac,Ramón Ceresd, Cristina Santos, Hybridization between multi-objective genetic algorithm and support vector machine for feature selection in walker-assisted gait , *computer methods and programs in biomedicine, Elsevier, 113 (2014) pp.736–748*

[7] H. Hannah Inbarania, Ahmad Taher Azarb, G. Jothi, Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis, *computer methods and programs in biomedicine, Elsevier, 113 (2014) pp.175–185*

[8] Wen-PingLi , JingYang , Jian-PeiZhang , Uncertain canonical correlation analysis for multi-view feature extraction from uncertain data streams, *Neuro computing, Elsevier, 149 (2015) pp.1337–1347*

[9] Marina E. Plissiti, Christophoros Nikou, Antonia Charchanti, Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images, *Pattern*

Recognition Letters, Elsevier 32 (2011) pp. 838–853

[10] S. Sasikala, S. Appavu alias Balamurugan, S. Geetha, Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set, *Applied Computing and Informatics (2014)*

[11] Gang Chen, JinChen, A novel wrapper method for feature selection and its applications, *Neuro computing, Elsevier 159(2015) pp. 219–226*

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

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

[14] Arvind R. Yadav, R.S. Anand, M.L. Dewal, Sangeeta Gupta, Multi resolution local binary pattern variants based texture feature extraction techniques for efficient classification of microscopic images of hardwood species, *Applied Soft Computing, Elsevier 32 (2015) pp. 101–112*

[15] YangCong, ShuaiWang , JiLiu , JunCao , YunshengYang , JieboLuo Deep sparse feature selection for computer aided endoscopy diagnosis, *Pattern Recognition, Elsevier, 48 (2015) pp. 907–917*

[16] Fengyi Song, XiaoyangTan, XueLiu, Songcan Chen Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients, *Pattern Recognition 47 (2014) 2825–2838*

[17] Mohamed Bennasar, YuliaHicks, Rossitza Setchi, Feature selection using Joint Mutual Information Maximisation, *Expert Systems With Applications, Elsevier 42 (2015) 8520–8532*

[18] B.Ashok , P.Aruna, Prognosis of Cervical Cancer using k-Nearest Neighbor Classification, *International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 9 (2015) 6676-6680*

[19] B.Ashok, Dr.P.Aruna, Analysis of Image Segmentation Methods for Detection of Cervical Cancer, *International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.75 (2015), 81-85*