IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 11-18
www.iosrjournals.org

COMPARISON OF CLASSIFICATION TECHNIQUES FOR PAP SMEAR DIAGNOSIS

P. Soorya praba[1], R.Priya[2]
[1]*PG Student,* [2]*Associate Professor*
Department of Computer Science and Engineering, Annamalai University, India

ABSTRACT:- *The Classification for Pap smear Diagnosis aims at classifying the Pap smear cells whether it is affected or not. The term "Pap-Smear" refers to samples of human cells stained by the so-called Papanicolaou method. The Papanicolaou method is a medical procedure to detect pre-cancerous cells in the uterine cervix. The median filter is used to remove the noises in the cell, and then the features are extracted using gray level co-occurrence matrix technique. k-NN classifier, Baye's classifier and ANN classifiers are used for the classification problem.. The classified cells are normal and abnormal.*

Keywords:- *Bayesian classifier**,** Genetic Algorithms, Nearest Neighbor based Classifiers, Neural Network, Pap-Smear Classification.*

I. INTRODUCTION

The term "Pap-Smear" refers to samples of human cells stained by the so-called Papanicolaou method. The Papanicolaou method is a medical procedure to detect pre-cancerous cells in the uterine cervix. A Pap smear is a microscopic examination of cells scraped from the opening of the cervix. The cervix is the lower part of the uterus that opens at the top of the vagina. Pap smear is a screening test for cervical cancer. Cervical cancer is malignant neoplasm of the cervix uteri or cervical area. It may present with vaginal bleeding, but symptoms may be absent until the cancer is in its advanced stages. Treatment consists of surgery in early stages and chemotherapy and radiotherapy in advanced stages of the disease. Pap smear screening can identify potentially precancerous changes. Treatment of high grade changes can prevent the development of cancer. In developed countries, the widespread use of cervical screening programs has reduced the incidence of invasive cervical cancer by 50% or more. The median filter method is used to remove the noises in the cell, and then the features are extracted using GLCM (gray level co-occurrence matrix) technique. k-NN classifier, Baye's classifier, and Artificial Neural Network classifiers are used for the classification problem.

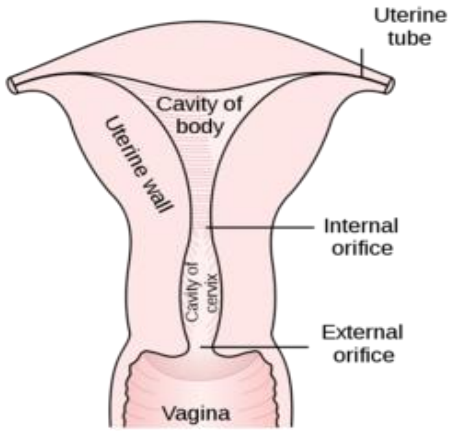

Fig.1 shows the cervix in relation to upper part of vagina and posterior portion of uterus

1 PROPOSED SYSTEM

In our work, we have considered two types of cells. One is normal cell and another one is abnormal cell.

2.1 Normal:

Our body is made up of billions of tiny cells that can only be seen under a microscope. These cells are grouped together to make up the tissues and organs of our bodies. They are a bit like building blocks.

2.2 Abnormal:

Malignant tumours are made up of cancer cells. They

- Usually grow faster than benign tumours
- Spread into and destroy surrounding tissues
- Spread to other parts of the body.

a) Normal cell **b)Abnormal cell**

Fig.2 Classified cells

In this project, classification of the pap smear cell is presented. The features are being extracted from the input images. The cell is classified in two different kinds, they are normal, benign, and malignant using k-Nearest Neighbor classifier. Finally we classify the pap smear cells.

The three different modules for the hybrid intelligent scheme for pap smear diagnosis are as follows,

1. Preprocessing
2. Feature Extraction
3. Classification
    - k – NN Classifier
    - Baye's Classifier
    - ANN Classifier

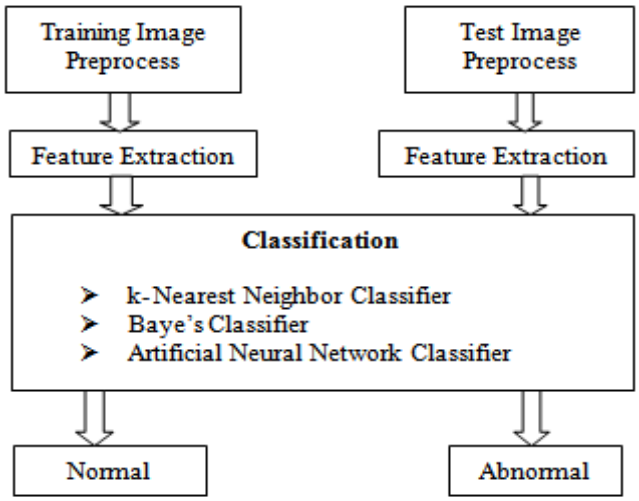

Fig.3 Block Diagram of the Proposed System

2.3 Preprocessing

In this module if any noise occurred in our input image, then such noises are removed. Noise is nothing but any undesired information that contaminates an image. These noises are removed using some type of filters. Here we are using Median filter to remove the noises.

2.3.1Median Filtering

We have seen that smoothing filters reduce noise. However, the underlying assumption is that the neighboring pixels represent additional samples of the same value as the reference pixel, i.e. they represent the

same feature. At edges, this is clearly not true, and blurring of features results. There are also nonlinear neighborhood operations that can be performed for the purpose of noise reduction that can do a better job of preserving edges than simple smoothing filters.
In the median filtering operation, the pixel values in the neighborhood window are ranked according to intensity, and the middle value becomes the output value for the pixel under evaluation. Median filtering does not shift boundaries, as can happen with conventional smoothing filters. Since the median is less sensitive than the mean to extreme values, those extreme values are more effectively removed.

2.4 Feature Extraction

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features also named features vector. Transforming the input data into the set of features is called *feature extraction*. The features provide the characteristics of the input type to the classifier by considering the description of the relevant properties of the image into a feature space. If the features extracted are carefully chosen, it is expected that they will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. We are using GLCM to extract the features.

2.4.1 GLCM

A gray level co-occurrence matrix (GLCM) contains information about the positions of pixels having similar gray level values. A co-occurrence matrix is a two-dimensional array, **P**, in which both the rows and the columns represent a set of possible image values. A GLCM **Pd** [i, j] is defined by first specifying a displacement vector **d**= (dx, dy) and counting all pairs of pixels separated by **d** having gray levels i and j.

2.4.2 Texture Features

1. Contrast

Contrast is a measure of the local variations present in an image.

$$C(k,n) = \sum_i \sum_j (i-j)^k \, P_d[i,j]^n \tag{1}$$

where,
k and n are the local variations, i and j are the pixel values, P_d is the co occurrence matrix. If there is a large amount of variation in an image the **P[i,j]**'s will be concentrated away from the main diagonal and contrast will be high (**typically k=2, n=1**).

2. Correlation

Correlation is a measure of image linearity.

$$C_c = \frac{\sum_i \sum_j [ijP_d[i,j]] - \mu_i \mu_j}{\sigma_i \sigma_j} \tag{2}$$

where,

$$\mu_i = \sum iP_d[i,j]$$
$$\sigma_i^2 = \sum i^2 P_d[i,j] - \mu_i^2$$

μ_i and σ_i are the mean and standard deviation, i and j are the pixel values of the function P_d(i,j). Correlation will be high if an image contains a considerable amount of linear structure.

3. Cluster Prominence

$$PROM = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \{i + j - \mu_x - \mu_y\}^4 \times P(i,j) \tag{3}$$

where,

i and j are the pixel values of the function P(i,j), G is gray tone, μ_x and μ_y are the mean values of P_x and P_y.

4. Cluster Shade

$$SHADE = \sum_{i=0}^{2G-2} (i - 2\mu)^3 \, H_s(i|\Delta x, \Delta y) \tag{4}$$

Where

$$\mu = \frac{1}{2} \sum_{i=0}^{2G-2} iH_s \, (i|\Delta x, \Delta y)$$

$\Delta y \; and \; \Delta y$ are the pixel distance occur within a given neighborhood, μ is the mean value, G- gray tone, H_s is the normalized sum histogram.

5.Dissimilarity

$$D = \sum_{i,j=1}^{G} C_{i,j} \, |i - j| \tag{5}$$

where,
i and j are the spatial coordinates of the correlation function $C_{i,j}$.

6. Energy

One approach to generate texture features is to use local kernels to detect various types of texture. After the convolution with the specified kernel, the texture energy measure (TEM) is computed by summing the absolute values in a local neighborhood:
If **n** kernels are applied, the result is an **n**-dimensional feature vector at each pixel of the image being analyzed.

$$L_e = \sum_{i=1}^{m} \sum_{j=1}^{n} |C(i,j)| \tag{6}$$

where,
i and j are the pixel values, m and n are the local kernels to detect various types of texture, C(i,j) is the absolute values of coefficient function.

7. Entropy

Entropy is a measure of information content. It measures the randomness of intensity distribution.

$$C_e = -\sum_{i} \sum_{j} P_d[i,j] \ln P_d[i,j] \tag{7}$$

where,
C_e= measure the entropy, N is the number of gray levels, i, j are the spatial coordinates of the function $P_{i,j}$. Such a matrix corresponds to an image in which there are no preferred gray level pairs for the distance vector d.

8. Homogeneity

A homogeneous image will result in a co-occurrence matrix with a combination of high and low P[i,j]'s.

$$C_h = \sum_{i} \sum_{j} \frac{P_d[i,j]}{1 + |i - j|} \tag{8}$$

where,
i and j are the pixel values of the co-occurrence matrix function P_d[i,j], the **range of gray levels** is small the **P[i,j]** will tend to be clustered around the main diagonal.

9. Maximum Probability

This is simply the largest entry in the matrix, and corresponds to the strongest response. This could be the maximum in any of the matrices or the maximum overall.

$$C_m = max P_d[i.j] \tag{9}$$

where,
C_m is the maximum probability, i and j are the pixel values of the co-occurrence matrix function $P_d[i,j]$.

10. Sum of Squares, Variance

$$VARIANCE = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(i-\mu)^2 P(i,j) \tag{10}$$

where,
i and j are the pixel values of the function P(i,j), G is gray tone, μ is the mean value. This feature puts relatively high weights on the elements that differ from the average value of P(i, j).

11. Autocorrelation

This function evaluates the linear spatial relationships between primitives. The set of autocorrelation coefficients shown below are used as texture features:

$$C(p,q) = \frac{MN}{(M-p)(N-q)} \frac{\sum_{i=1}^{M-p}\sum_{j=1}^{N-q} f(i,j)f(i+p,j+q)}{\sum_{i=1}^{M}\sum_{j=1}^{N} f^2(i,j)} \tag{11}$$

where p, q is the positional difference in the i, j direction, and M, N are image dimensions.

12. Local Homogeneity, Inverse Difference Moment (IDM)

$$IDM = \sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{1}{1+(i-j)^2}P(i,j) \tag{12}$$

IDM is also influenced by the homogeneity of the image. Because of the weighting factor IDM will get small contributions from inhomogeneous areas.

2.5 Classification

In this classification problem, k-NN classifier, Baye's classifier and ANN classifiers.

2.5.1 Nearest Neighbor Classifier

Initially, the classic 1-Nearest Neighbor (1-nn) method is used. The 1-nn works as follows: In each iteration of the feature selection algorithm, a number of features are activated. For each sample of the test set its Euclidean Distance from each sample of the training set is calculated. The Euclidean Distance is calculated as follows:

$$D_{ij} = \sqrt{\sum_{l-1}^{d}\left|x_{il} - x_{jl}\right|^2} \tag{13}$$

where Dij is the distance between the test sample xil and the training sample xjl, and l = 1, …, d is the number of activated features in each iteration. With this procedure the nearest sample from the training set is calculated. Thus, each test sample is classified in the same class that its nearest sample from the training set belongs. The previous approach may be extended to the k-Nearest Neighbor (k-NN) method, where we examine the k-nearest samples from the training set and, then, classify the test sample by using a voting scheme. The most common way is to choose the most representative class in the training set. Thus, the k-NN method makes a decision based on the majority class membership among the k nearest neighbors of an unknown sample. In other words every member among the k nearest has an equal percentage in the vote.

A nearest-neighbor classification object, where both distance metric ("nearest") and number of neighbors can be altered. The object classifies new observations using the predict method. The object contains the data used for training, so can compute resubstitution predictions. The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data. Nearest neighbor rules in effect implicitly compute the decision boundary. It is also possible to compute the decision boundary explicitly, and to do so efficiently, so that the computational complexity is a function of the boundary complexity. Nearest

neighbor problem has been extensively studied in the field of Computational geometry under the name closest pair of point's problem.

2.5.2 Baye's Classifier

Gaussian baye's classifier is a baye's classifier for data input classes having Gaussian distribution. The classifier learns from training data and estimates the posterior probabilities of the classes given particular instance of the features using baye's theorem. Prediction of the class is determined by identifying the class with the highest posterior probability. The major advantage of the baye's classifier is its short computational time for training since it requires a relatively small amount of training data to estimate the parameters for classification. Baye's classifier is also robust to missing values because these values are simply ignored in computing probabilities and thus have no impact on the final decision. The baye's classifier is a traditional statistics-based classifier that analyzes discriminant functions.

$$P(h|D)=\frac{P(D|h)\,P(h)}{P(D)} \tag{14}$$

Where,

D - feature values of feature images,

P(h|D) - probability of diseases,

P(D|h) - the posterior probability function of normal and diseased category,

P(D) – the prior probability of training data,

P(h) – the prior probability of hypothesis.

2.5.3 Artificial Neural Network

The structure of a neural network is formed by an "input" layer, one or more "hidden" layers, and the "output" layer. The number of neurons in a layer and the number of layers depends strongly on the complexity of the system studied. Therefore, the optimal network architecture must be determined. The general scheme of a typical three-layered ANN architecture is given in Fig.4.

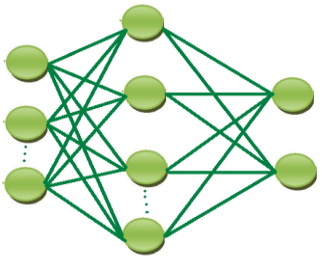

Fig.4 Structure of ANN

Input Layers are loaded the data (extracted features). Create the network. Network has two output neurons are normal and abnormal Pap smear cell. Each output neuron represents the class whether it is normal or abnormal. Setup the division of data. Input vectors and target vectors are randomly divided into 70% - training, 15% - validation, 15% - testing. We training and testing the datasets. Finally classify the cell.

II. EXPERIMENTAL RESULTS

2.1 Performance Calculation

Sensitivity and specificity are the statistical measure of the performance of a binary classification test, also known in statistics or classification functions, sensitivity measures the proportion of actual positives which are correctly identified. Specificity measures the proportion of negatives which are correctly identified.

- TP=True Positive is correctively classified as positive pixels.
- FN=False Negative is incorrectly classified as positive pixels.
- FP=False Positive is incorrectly classified as negative pixels.
- TN=True Negative is incorrectly classified as negative pixels.

True Positive: Pap smear affected images are correctly identified as Pap smear.
False Positive: Normal image incorrectly identified as Pap smear.
True Negative: Normal image correctly identified as Normal.
False Negative: Pap smear image incorrectly identified as Normal.

CLASSIFIERS	SENSITIVITY	SPECIFICITY	ACCURACY
k-NN	92	80	85
BAYE'S	88.89	60	73.68
NEURAL NETWORK	100	90	94.74

Table.1 Performance measures

3.1.1 Sensitivity:

Sensitivity measures the proportion of actual positives which are correctly identified.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{15}$$

3.1.2 Specificity:

Specificity measures the proportion of negatives which are correctly identified.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{16}$$

3.1.3 Accuracy:

Accuracy of the measurement system is the degree of closeness of measurement of the quantity to the quantities of actual (true) value. Accuracy is also used as the statistical measures of how well a binary classification test correctly identifies or exclude conditions. Accuracy is the proportion of true result in the population.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{17}$$

III. CONCLUSION AND FUTURE WORK

In this project, a classification algorithm is proposed for solving the Pap-smear cell classification problem. In the pre-processing method, the median filter is used to remove the noises in the cell. Then the features are extracted using GLCM technique, the features such as contrast, correlation, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity, maximum probability, sum of squares and variance, autocorrelation, and inverse difference moment.. The k-Nearest Neighbor classifier, Baye's classifier and Artificial Neural Network classifiers are used to classify the Pap smear cells, the three types of classified cells are normal, benign and malignant. The performance of the proposed algorithm is tested using data sets of Pap-smear cells. The obtained results indicate the high performance of the proposed algorithm (Artificial Neural Network) in searching for a reduced set of features with high accuracy and in achieving excellent classification of Pap-smear cells. In the future work, intended to be focused in using different classifiers and different techniques for extraction.

REFERENCES

[1]. Aha D.W. and Bankert R.L., 1996, "A Comparative evaluation of sequential feature selection algorithms". In Artificial Intelligence and Statistics, Fisher D. and Lenx J.-H. (Eds.), Springer-Verlag, New York.
[2]. Al-Ani A., 2005a, "Feature subset selection using ant colony optimization". International Journal of Computational Intelligence, 2(1), pp. 53-58.
[3]. Al-Ani A., 2005b, "Ant colony optimization for feature subset selection". Transactions on Engineering, Computing and Technology, 4, pp. 35-38.
[4]. Byriel, J., 1999, "Neuro-fuzzy classification of cells in cervical smears". Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation.
[5]. Cantu-Paz E., 2004, "Feature subset selection, class separability, and genetic algorithms". Genetic and Evolutionary Computation Conference, pp. 959-970.
[6]. Cantu-Paz E., Newsam S. and Kamath C., 2004, "Feature selection in scientific application". Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 788-793.
[7]. Dorigo M. and Stützle T., 2004, Ant Colony Optimization. A Bradford Book, The MIT Press Cambridge, Massachusetts, London, England.
[8]. Duda R.O. and Hart P. E., 1973, Pattern Classification and Scene Analysis. John Wiley and Sons, New York,.
[9]. Goldberg, D. E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company, INC, Massachussets.
[10]. Holland, J. H., 1975, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI.
[11]. Jain A. and Zongker D., 1997, "Feature selection: Evaluation, application, and small sample performance". IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, pp. 153-158.
[12]. Jantzen, J., Norup, J., Dounias, G. and Bjerregaard B., 2006, "Pap-smear benchmark data for pattern classification". (submitted).
[13]. Kira K. and Rendell L., 1992, "A practical approach to feature selection". Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, pp. 249-256.
[14]. Kohavi R. and John G., 1997, "Wrappers for feature subset selection". Artificial Intelligence, 97, pp. 273-324.
[15]. Lopez F.G., Torres M.G., Batista B.M., Perez J.A.M. and Moreno-Vega. J.M., 2006, "Solving feature subset selection problem by a parallel scatter search". European Journal of Operational Research, 169, pp. 477-489.
[16]. Marinakis Y., Migdalas, A., and P. M. Pardalos, 2005, "A Hybrid Genetic-GRASP algorithm Using Langrangean Relaxation for the Traveling Salesman Problem", Journal of Combinatorial Optimization, 10, pp. 311-326.
[17]. Martin, E., 2003, "Pap-smear classification", Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation,.
[18]. Narendra P.M. and Fukunaga K., 1977, "A branch and bound algorithm for feature subset selection". IEEE Transactions on Computers, 26(9), pp. 917-922.
[19]. Norup, J., 2005, "Classification of pap-smear data by transductive neuro-fuzzy methods", Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation,.
[20]. Parpinelli R.S., Lopes, H.S., and Freitas, A.A., 2002, "An ant colony algorithm for classification rule discovery". In Data mining: A heuristic approach, Abbas H., Sarker R. and Newton C. (Eds.), London, UK: Idea group publishing, pp. 191–208.