FEATURE EXTRACTION OF INTRADUCTAL BREAST IMAGES USING GMM

Ms.G.Prieyadharsini[1], Prof.P.Tamije Selvy[2] and Dr.V.Palanisamy[#]

[1]PG Student, [2]Assistant Professor(SG)
[1,2]Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, India.
[#]Principal, Info Institute of Engineering, Coimbatore, India.

[1]gprieyadharsini@gmail.com, [2]tamijeselvy@gmail.com

ABSTRACT

Intraductal Carcinoma is a noninvasive condition in which abnormal cells are found in the lining of a breast duct. The abnormal cells have not spread outside the duct to other tissues in the breast. In some cases, Intraductal Carcinoma may become invasive cancer and spread to other tissues, although it is not known at this time how to predict which lesions will become invasive. Intraductal cancer is the most common type of breast cancer in women. Memory Intraductal includes 3-types of cancer: Usual Ductal Hyperplasia (UDH), Atypical Ductal Hyperplasia (ADH), and Ductal Carcinoma in Situ (DCIS). So the system of detecting the breast microscopic tissue of UDH, ADH, DCIS is proposed. The current standard of care is to perform percutaneous needle biopsies for diagnosis of palpable and image-detected breast abnormalities. UDH is considered benign and patients diagnosed UDH undergo routine follow-up, whereas ADH and DCIS are considered actionable and patients diagnosed with these two subtypes get additional surgical procedures. The system classify the tissue based on the quantitative feature derived from the images. The statistical features are obtained. The approach makes use of preprocessing, Cell region segmentation, Individual cell segmentation, Feature extraction technique for the detection of cancer.

KEYWORDS

Intraductal Carcinoma, percutaneous, Cell Segmentation.

1. INTRODUCTION

Medical imaging is one of the fastest growing areas within medicine at present, both in the clinical setting in hospitals. Medical imaging is the technique and process used to create images of the human body for clinical purposes or medical science. Medical imaging is often perceived to designate the set of techniques that noninvasively produce images of the internal aspect of the body. Medical imaging can be seen as the solution of mathematical inverse problems. This means that cause is inferred from effect. This is very important to help improve the diagnosis, prevention and treatment of the diseases. Medical imaging is a part of biological imaging and incorporates radiology, nuclear medicine, investigative radiological sciences, endoscopy, thermography, medical photography and microscopy.

1.1 Background

1.2

The continuum of intraductal breast lesions, which encompasses the usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), and ductal carcinoma in situ (DCIS), are a group of cytologically and architecturally diverse proliferations, typically originating from the terminal

duct-lobular unit and confined to the mammary duct lobular system. These lesions are highly significant as they are associated with an increased risk of subsequent development of invasive breast carcinoma, albeit in greatly differing magnitudes. Clinical follow-up studies indicate that UDH, ADH, and DCIS are associated with 1.5, 4–5, and 8–10 times of increased risk respectively, compared to the reference population for invasive carcinoma.

Patients diagnosed UDH are advised to undergo routine follow-up, while those with ADH and DCIS are operated by excisional biopsy followed by possible other surgical and therapeutic procedures. The pathology diagnoses are typically made according to a set of criteria defined by the World Health Organization (WHO), using formalin fixed paraffin embedded tissue specimens, which are stained with a mixture of hematoxylin/eosin (H&E), no single criterion is absolute[4]. Thus, subjective assessment and weighing the relative importance of each criterion are performed to categorize the lesions. The proposed system applies segmentation and feature extraction techniques for detection of cancer.

1.3 Breast Lesions

A lesion is an area which is an abnormality or alteration in the tissue's integrity. Lesions can occur in any area of the body consisting of soft tissue, commonly found on the skin. There are numerous types of lesions with different naming classifications. When this lesion develops in the breast tissues, they are referred to as breast lesions. Breast lesions usually come in the form of lumps or swellings in or around the breast area, and they are frequently felt during a self breast examination or when examined by a physician. Some breast lesions, however, may be present but not felt. These are called non-palpable lesions, and they are mostly detected during a screening mammogram test, which is more like an x-ray of the breast.

The normal breasts have various types of tissues with different consistencies. One type of tissue found in the breasts is the glandular tissue, which is nodular and firm to the touch. Breasts also have fats that are generally soft to the touch. It is normal for the breast tissues to undergo changes such as lumpiness or tenderness, especially during the menstrual cycle. Most of these breast changes normally occur in response to hormonal changes going on in the body.

Breast Lesions[6] (malignant breast neoplasm) is cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. Cancers originating from ducts are known as ductal carcinomas; those originating from lobules are known as lobular carcinomas. Breast cancer is a disease of humans and while the overwhelming majority of cases in humans are women. Size, stage, rate of growth, and other characteristics of the tumour determine the kinds of treatment. Treatment may include surgery, drugs, radiation and/or immunotherapy. Surgical removal of the tumour provides the single largest benefit, with surgery alone being capable of producing a cure in many cases.

2. RELATED WORK

J. Rozai et al. [4] focuses on "borderline lesions" of the breast. ADH is the most popular and classical borderline breast lesion. It is defined as the atypical intraductal epithelial proliferation, which mimics low-grade ductal carcinoma in situ (DCIS), and they are mostly small. The significance of ADH is its relative risk for developing invasive carcinomas on both breasts. Some breast carcinomas with peripheral ADH associations are found. In those cases, the area of peripheral ADH are recommended to be removed together with carcinoma, as they may be connected to the main lesion through the duct profiles.

P. L. Fitzgibbons et al. [2] states that tumor size as a prognostic variable in cases of invasive carcinoma is robust. It is used to measure the various clinical estimates and mammograms. Tumor size is directly related to an increasing probability of regional metastasis, increasing average number of auxiliary lymph nodes and probability of recurrence and death. The favorable prognosis of non palpable invasive carcinoma is relative to palpable ones. Precise assessment of tumor size is necessary to properly stratify patients, particularly since screening mammography has resulted in a steadily increasing proportion A.P.Dempster, N. M. Laird, et al. [1] proposed a general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. Since each iteration of the algorithm consist of an expectation step followed by a maximization step, it is called as EM algorithm. The EM algorithm is remarkable process because of its simplicity and generality of the associated theory.

L. Vincent and P. Soille, et al.[5] states that Watersheds are one of the classics in the field of topography, a gray-level image is considered a topographic relief where the gray level of a pixel is interpreted as its elevation. The water flows along a topographic relief following a certain descending path to eventually reach a catchment basin. Blobs in the image can be separated using this concept by identifying the limits of adjacent catchment basins and then separating them. The lines separating catchment basins are called watersheds.

D. Page and W. Dupont, et al[6] states that breast imaging have made a positive impact on breast cancer screening and detection. The growing use of image-detected biopsies has led to increased diagnosis of ductal carcinoma in situ and high-risk proliferative breast lesions. This progress, has created a challenge for pathologists. In lieu of the fact that these entities are difficult to diagnose even in tissue sections taken from surgically excised lesions, pathologist are now expected to diagnose them in small and often fragmented tissue/cellular samples obtained from image-guided biopsies. Some proliferative lesions are associated with an increased risk of finding neighboring malignant.

3. PROPOSED SCHEME

The proposed system applies preprocessing, Cell region segmentation, Individual cell segmentation, Feature extraction technique for the detection of cancer. The first step of preprocessing involves the min-max normalization preprocessing. Three different lesion subtypes are used: Clustering algorithm is used to identify region of cells in the H&E stained breast microscopic tissue. This was followed by a watershed-based algorithm which identifies individual cells. The segmented cells are used to derive size, shape and intensity based feature characterizing each ROI. Segmentation is done using cell region segmentation and individual cell segmentation. Feature extraction is implemented using ROI.

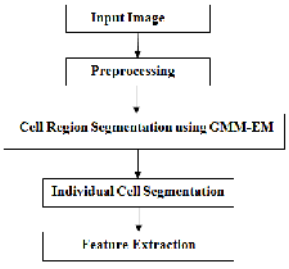


Figure 1. Overall Flow of the System.

3.1 Preprocessing

Preprocessing can be applied easily. It improves the effectiveness and performance. Min-max normalization is used for preprocessing. Min-Max normalization is the process of taking data measured in its engineering units and transforming it to a value between 0.0 and 1.0. The lowest (min) value is set to 0.0 and the highest (max) value is set to 1.0. It provides an easy way to compare values that are measured using different scales or different units of measure.

3.2 Segmentation

The purpose of image segmentation is to partition an image into meaningful regions with respect to a particular application .It is based on measurements taken from the image and might be greylevel, colour, texture, depth or motion. It is to partition a image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. Cell region segmentation and Individual cell segmentation are used to segment the input breast lesion.

3.2.1 Cell Region Segmentation

Cell segmentation would be the first step toward automated analysis of histopathological slides. This is implemented in two steps. In the first step, cell regions are segmented by clustering the pixel data and in the second step segmented cell regions are further processed by a watershed-based segmentation algorithm to identify individual cells. The cell region segmentation performs the following steps:

1) ROI images are converted into RGB color space then to L a*b*.
2) La*b* color space also separates the luminance and the chrominance information such that: L channel corresponds to illumination and a*and b* channels correspond to color-opponent dimensions.

The equation to convert RGB to La*b* color space are as follows:

L=(116*var_Y)-16;a=500*(var_X-var_Y); b=200*(var_Y- var_Z)

Segmentation performs maximum likelihood estimation of gaussian mixture model by using expectation algorithm [1]

EM Algorithm

Expectation step (E-step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of given under the current estimate of the parameter:

$$Q(\theta|\theta^{(t)})=E_{Z|X,\theta}{}^{(t)}[\log(\theta;X,Z)]$$

Maximization step (M-step): The parameter that maximizes this quantity:

$$\theta^{(t+1)}=\arg\max_{\theta} Q(\theta|\theta^{(t)})$$

The expectation maximization (EM)[1] algorithm is implemented using the a*, b* channels to estimate the parameters of the GMM model. The resulting mixture distribution is used to classify pixels into four categories. Those classified into the cellular component are further

clustered in the L channel by dynamic thresholding [7] to eliminate blue–purple pixels with relatively less luminance.

3.2.2 Individual Cell Segmentation

Segmentation maps of cell regions obtained are converted to graylevel images before they are used in this stage. Since most segmented regions contain multiple overlapping cells with cells only vaguely defined due to the presence of holes inside them, connected components in these images do not necessarily represent individual cells. These images are first preprocessed using hole filling and cleaning steps suggested in[8]. Overlapped cells result in blobs in the segmentation map. To separate these blobs properly so as to identify individual cells, we used a watershed algorithm [5] based on immersion simulations. The watersheds algorithm performs the following steps:

1) RGB image obtained are converted to gray-level images.
2) A gray-level image is considered a topographic relief where the gray level of a pixel is interpreted.
3) The water flows along a topographic relief following a certain descending path to eventually reach a catchment basin.

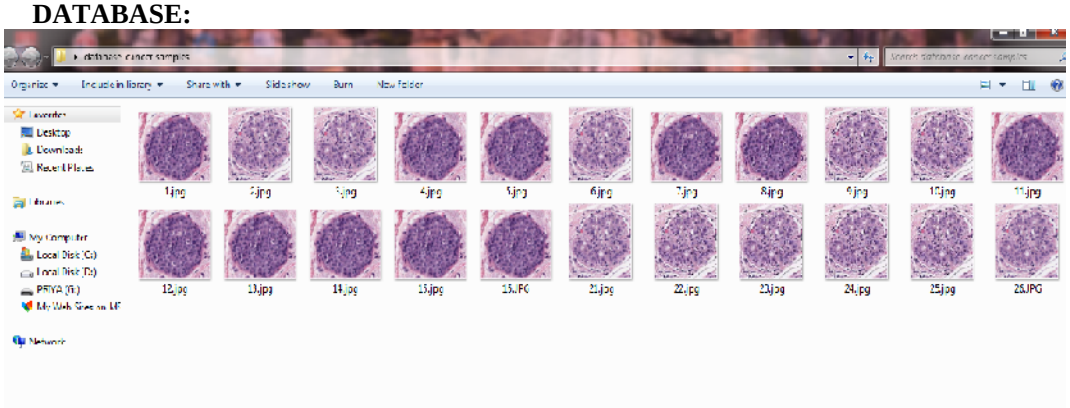
3.3 Feature Extraction

Feature extraction and selection methods are to obtain the most relevant information from the original data and represent that information in a lower dimensionality space. When the cost of the acquisition and manipulation of all the measurements is high we must make a selection of features. The goal is to select, among all the available features, those that will perform better. Example: Features that should be used for classifying a student as a good or bad one. The available features for student classification are :marks, height, sex, weight, IQ. Feature selection would choose marks and IQ and would discard height, weight and sex. The feature extraction performs the following steps:

1) The perimeter, the ratio of major to minor axis, and the mean of the gray-level intensity are computed.
2) For each connected component identified in an ROI.
3) Statistical features involving the mean, standard deviation, median, and mode are computed to obtain features at the ROI level.
4) Thus, each ROI is characterized by a total of 12 features (3 × 4).

The database which is used for detection of cancer

DATABASE:



4. EXPERIMENTAL RESULTS

Fig. 2 shows the stain image of breast lesion. Lesions can be seen in the stain image. Breast Lesions is cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. Cancers originating from ducts are known as ductal carcinomas. Fig. 3 shows the preprocessed image using the min-max normalization method. In MIN-MAX normalization the values lies between -1 and 0.It avoids the numerical problems.

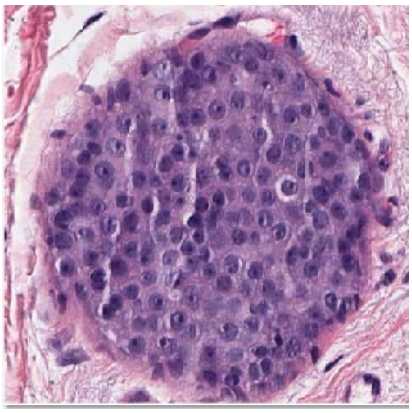


Figure 2. Stain Image

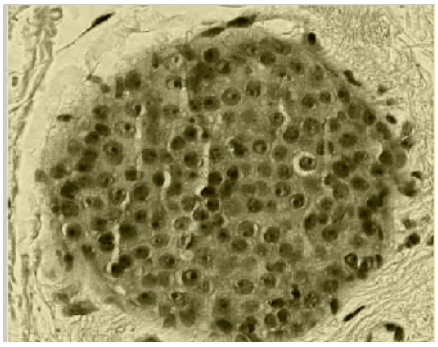


Figure 3. Min-Max Preprocessed Image

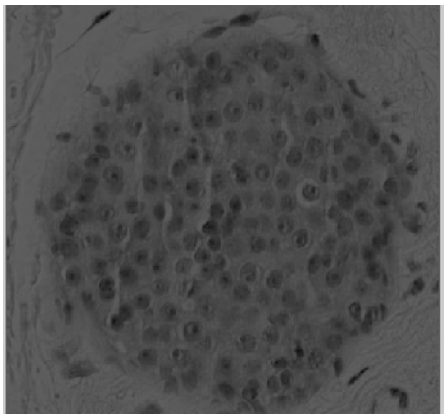


Figure 4. L-Channel Image

Fig. 4,5,6 shows the La*b* image. The RGB image of breast lesion is converted to L a*b* images and from that L, a, b-channels are extracted.

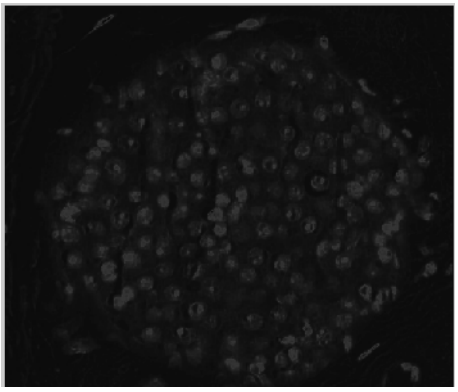


Figure 6. B-Channel Image

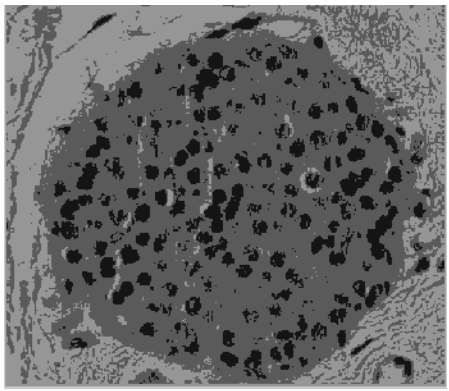


Figure 7. L-Segmented Cell Regions

Figure 8. A-Segmented Cell Regions

The results were over segmented by both methods; by using cell region segmentation and individual cell segmentation segmented cells were detected. Fig. 7, 8,9 shows the segmented cell regions of L a*b* image.

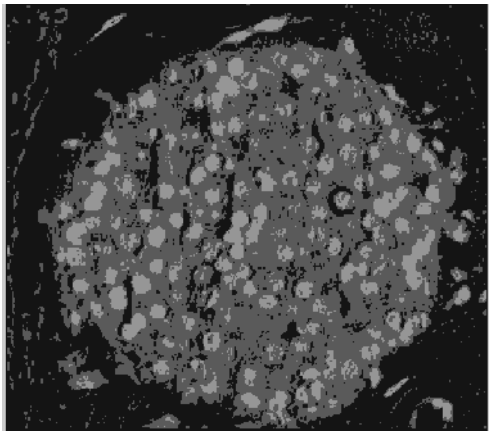


Figure 9. B-Segmented Cell Regions

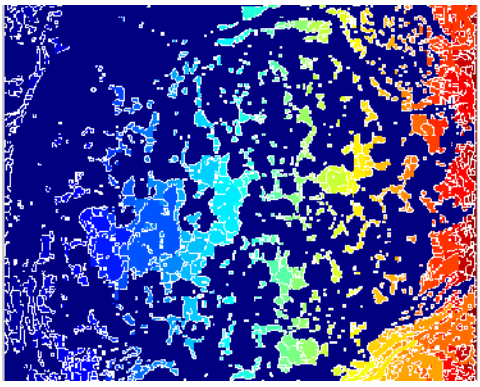


Figure 10. Individual Segmented Cells

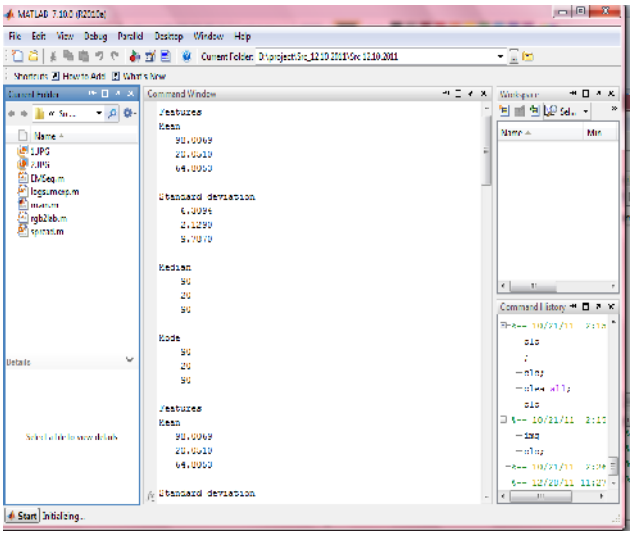


Figure 11. Statistical Features

Fig. 10 shows the segmented cells which are detected using individual cell segmentation. Fig. 11 shows the statistical features which are being extracted.

Table 1. Accuracy based on precision and Recall

Accuracy	Precision	Recall
Computerized System	87.9	10.6
Pathologist's Prediction	82.6	15.5

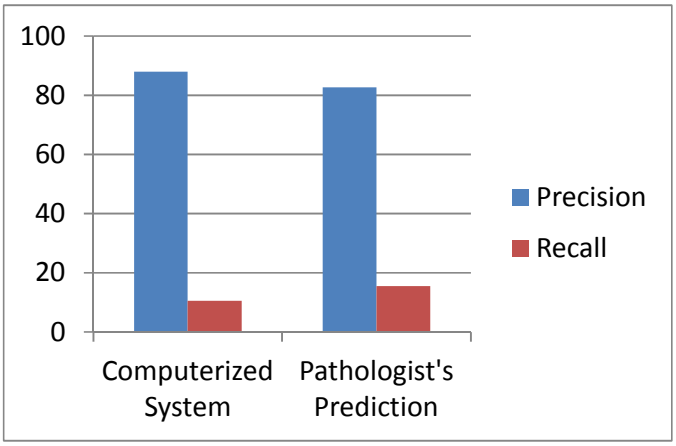


Figure 12. Accuracy of Computerised system and pathologist's prediction

Table 1 represents the accuracy of segmentation model based on precision and recall. All segmented cells obtained from different image segmentation models are ranked as shown in Fig. 12. An overall accuracy of 87.9% precision and 10.6% recall are achieved on the entire test data. The test accuracy of 82.6% precision and15.5% recall are obtained with borderline cases only, when compared against the diagnostic accuracies of pathologists on the same set, indicates

that the system is highly competitive with the expert pathologists as a stand-alone diagnostic tool and has a great potential in improving diagnostic accuracy and reproducibility.

5. CONCLUSIONS

The proposed cell region segmentation, individual cell segmentation and feature extraction is used to identify the breast lesions. EM algorithm is used for cell segmentation. In this approach step of initialization is necessary to prevent settling down on a bad local maximum. Then the EM algorithm gets an opportunity to explore the parameter space and it may converge to a better maximum. Generally, the clustering-based initialization method provides a better final result for the EM algorithm than random initialization does, and it also contributes to the convergence speed. This will involve developing intermediate models to map image features onto descriptors pathologists use for classification. This new approach can help as an automated medical image analysis tool to validate our hypothesis in an accurate and specific manner. Future work includes implementation of MIL and SVM classifier for detecting the type of breast lesions. The MIL and SVM could further enhance the result, which could closely detecting the type of breast lesions

REFERENCES

[1] A.P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm.," J. R. Statist. Soc. SeriesB (Methodol.), vol. 39, no. 1, pp. 1–38, 1977.

[2] P. L. Fitzgibbons, D. L. Page, D. Weaver, A. D. Thor, D. C. Allred, G.M. Clark, S. G. Ruby, F. O'Malley, J. F. Simpson, J. L. Connolly, D.F. Hayes, S. B. Edge, A. Lichter, and S. J. Schnitt, "Prognostic factor in breast cancer," *Arch. Path ol. Lab. Med.*, vol. 124, no. 7, pp. 966–978, 2000.

[3] R. Jain, R. Mehta, R. Dmitrov, L. Larsson, P. Musto, K. Hodges, T. Ulbright, E. Hattab, N.Agaram, M. Idrees, and S. Badve, "A typical ductal hyperplasia at 25 years-interobserver and intraobserver variability," *Mod. Pathol.*, vol. 23, no. 1, suppl. 1, pp. 53A:abstr. 229, 2010.

[4] J. Rozai, "Borderline epithelial lesions of the breast," *Amer.* J. Surg. Pathol., vol. 15, no. 3, pp. 209–221, 1991.

[5] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," IEEE Pattern Anal.Mach. Intell., vol. 13, no. 6, pp. 583–98, Jun. 1991.

[6] D. Page, W. Dupont, L. Rogers, and M. Rados, "Borderline epithelial lesions of the breast," *Amer. J. Surg. Pathol.*, vol. 15, pp. 209–221,1991.

[7] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan.1979.

[8] S. Eddins, "Cell segmentation," (2006). [Online].
Available: http://blogs.mathworks.com/steve/2006/06/02/cell-segmentation/

Authors

Ms.G.Prieyadharsini has received Bachelor of Technology degree in Information Technology under Anna University, Chennai in 2010. She is currently pursuing Master of Engineering degree in Computer Science and Engineering under Anna University, Coimbatore, India. Her areas of interest are image processing and data mining.

Prof. P.Tamije Selvy received B.Tech (CSE), M.Tech (CSE) in 1996 and 1998 respectively from Pondicherry University. Since 1999, she has been working as faculty in reputed Engineering Colleges. At Present, she is working as Assistant Professor(SG) in the department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore. She is currently pursuing Ph.D under Anna University, Chennai. Her Research interests include Image Processing, Data Mining, Pattern Recognition and Artificial Intelligence.

Dr.V.Palanisamy received B.E (Electronics and communication), M.E (Communication Engineering) and PhD (Communication Engineering) in 1972, 1974 and 1987 respectively. Since 1974, he has been the faculty of Electronics and Communication Engineering and Served at various Government engineering colleges. At present, he is the principal at Info Institute of Engineering, Coimbatore. His research interest is in the Heuristic search methods for Optimization problems in various applications. He is a Senior member of ISTE, SIEEE, and CSI.