Feature Extraction Method for the Discovery of Breast Cancer Lacerations by Using Mixture Model and EM Algorithm

Sathish kumar.M, Dinesh.E, MohanRaj.T

Abstract —Breast cancer is a type of cancer begins from breast tissue, most generally from the inner lining of milk ducts or the lobules that supply the ducts with milk. Cancers invent from ducts are identified as ductal carcinomas, while those originating from lobules are known as lobular carcinomas. Breast cancer occurs in humans and other mammals. While the vast majority of human cases occur in women, male breast cancer can also occur. Intraductal Carcinoma is a noninvasive condition in which abnormal cells are found in the lining of a breast duct. The irregular cells have not spread outside the duct to other tissues in the breast. In some cases, Intraductal Carcinoma may become persistent 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 systems 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.

Index Terms— Intraductal Carcinoma, percutaneous, Cell Segmentation, computer-aided diagnosis, Histopathological image analysis, intraductal breast lesions, Breast cancer

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. It is often perceived to designate the set of techniques that noninvasively produce images of the internal aspect of the body. It 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.

- *M.Sathish kumar is currently pursuing master's degree program in computer science and engineering in Karpagam University, Coimbatore,Tamilnadu. India, E-mail: msathishkumar.moorthy@gmail.com*

- *E.Dinesh is currently pursuing masters degree program in computer science and engineering in Karpagam University, Coimbatore,Tamilnadu. India, E-mail: dineshelayaperumal@gmail.com*

- *T.Mohanraj is currently working as an assistant professor in the department of computer science and engineering in Karpagam University, Coimbatore,Tamilnadu. India, E-mail: mohanrajt.me@gmail.com*

1.1 Background

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 ematoxylin/eosin (H&E), no single criterion is absolute. 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.

1.2 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. 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; those originating from lobules are known as lobular carcinomas. 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.

Benign tumors:

Benign tumors are rarely a threat to life. This can be removed and usually don't grow back. It doesn't invade the tissues around them. Moreover don't spread to other parts of the body.

Malignant tumors:

Malignant tumors may be a threat to life. Often can be removed but sometimes grow back It can invade and damage nearby organs and tissue scan spread to other parts of the body.

Breast cancer cells can spread by breaking away from the original tumor. They enter blood vessels or lymph vessels, which branch into all the tissues of the body. The cancer cells may be found in lymph nodes near the breast. The cancer cells may attach to other tissues and grow to form new tumors that may damage those tissues. The spread of cancer is called metastasis. Breast MRI is the future of breast radiology. It is a powerful tool in detecting early cancers which are not even seen on mammograms. Because of the cost, it is used as a problem solving modality. It's the most sensitive tool available hence certain lesions seen on MRI may not be seen on other imaging modalities hence MRI-guided localization or a biopsy system are needed. Women at high risk can undergo screening MR mammography.

1.3 Related Work

Imaging techniques in X-Ray, MRI and Ultrasound diagnostics yield a great deal of information, which the radiologist has to analyze and evaluate comprehensively in a short time. CAD systems help scan digital images, e.g. from computed tomography, for typical appearances and to highlight conspicuous sections, such as possible diseases. CAD is a relatively young interdisciplinary technology combining elements of artificial intelligence and digital image processing with radiological image processing. A typical application is the detection of a tumor. For instance, some hospitals use CAD to support preventive medical check-ups in mammography (diagnosis of breast cancer), the detection of polyps in the colon, and lung cancer. The microscopic examination of tissue in order to study the manifestations of disease. Specifically, in clinical medicine, histopathology refers to the examination of a biopsy or surgical specimen by a pathologist, after the specimen has been processed and histological sections have been placed onto glass slides. In contrast, cytopathology examines free cells or tissue fragments. In order to train the system to perform multicategory classification, samples with reference standard from each subtype would be necessary. Since there is currently no known morphometric, immunohistochemical, or molecular features to distinguish ADH from low grade DCIS, such a reference standard could not be established for ADH and certain types of DCIS. Thus, the reference standard required for the training of the classifier for classifying UDH versus actionable lesions can be established with a reasonable effort, whereas the same cannot be said for the reference standard required for multicategory classification.

2. 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 and Feature extraction is implemented using ROI.

Goals and Objectives

Goals

- To perform diagnoses for the patient by encapsulate the patient information in the medical image.
- To extract the patient information from the segmentation of medical image.

Objectives

- To provide the patient information for diagnosing the actionable subtype in the medical image.
- To provide lower computational complexity and higher diagnosing capacity.

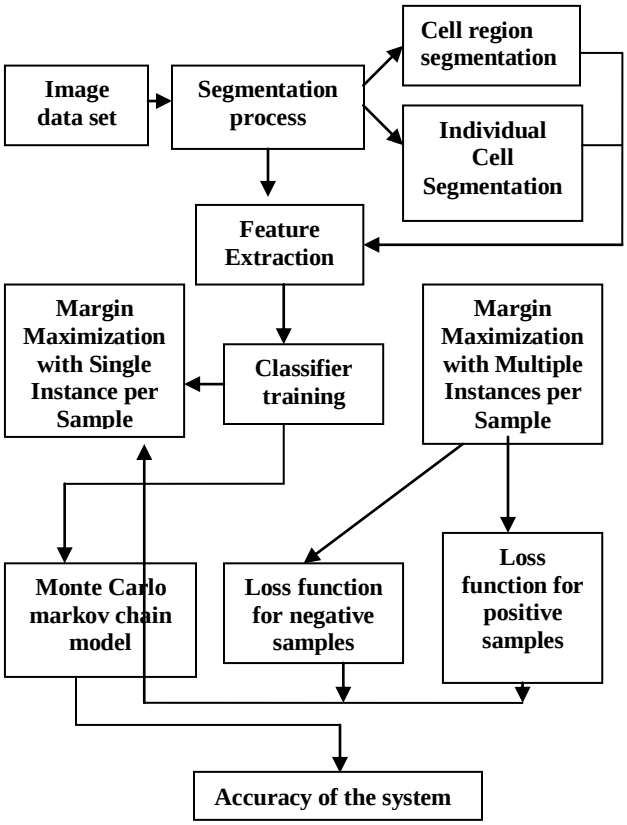

Fig.1 Architecture of Proposed Method

2.1 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.

2.2 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, color, texture, depth or motion. It is to partition an 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.

2.3 Proposed Algorithm:

Markov Chain Monte Carlo methods do not sample from the distribution of interest P(x) directly. Instead, they sample from a sequence of distributions that converges to P(x).In proposed system we describe statistical inference of neighbor-dependent models using a Markov chain Monte Carlo expectation maximization (MCMC-EM) algorithm. In the MCMC-EM algorithm, the high-dimensional integrals required in the EM algorithm are estimated using MCMC sampling. The MCMC sampler requires data sample from continuous time Markov process, conditional on the beginning and ending states and the paths of the neighboring models. These MCMC-EM clustered features are used for training and testing the binary classifier. For classification we use the extended version of MIL (multiple instance learning) classifier.

2.4 Advantages

The proposed clustering algorithm also resulted in better (i.e., smaller) entropy values than the GMM-EM algorithm in all cases. We can argue that the reason behind this performance is that the proposed algorithm does not need data for explicit estimation of the cluster parameters because it generates the parameters via update equations, whereas EM is highly data dependent in the calculation of the parameters.

3. METHODS

3.1 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 watershedbased 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. Segmentation performs maximum likelihood estimation of gaussian mixture model by using expectation algorithm.

3.2 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. 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 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).

3.4 Classification method:

Each slide contains multiple ROIs and a positive diagnosis is confirmed when at least one of the ROIs in the slide is identified as positive. For a negative diagnose, the pathologist has to rule out the possibility of each and every ROI being actionable. The objective here is to develop a classifier to optimize classification accuracy at the slide level. Traditional supervised training techniques which are trained to optimize classifier performance at the instance level yield suboptimal performance in this problem.

The parameters of the classifiers are selected from a designated set of six different C+ and C− values by considering all possible pairs and selecting the pair that optimizes the leave-one-slide-out (LOSO) cross-validation performance of the classifier. LOSO cross-validation splits the training dataset into k folds, where k is equivalent to the number of slides. At each stage, one slide is left out as validation data, i.e., all ROIs for that slide are removed from the training data as validation data, and the classifier is trained with the ROIs of the remaining k – 1 slide and tested on the ROIs of the left-out slide. This process is repeated until all k slides are used for validation and the probabilities of all ROIs being positive are obtained. The proposed system will make the most clinical impact when developed to address the binary classification of UDH versus actionable subtypes during the percutaneous biopsy stage. Clinical impact aside, the binary classification approach is more feasible from the system training perspective as well. GMM-EM easily trapped in local minima, and they are very sensitive to initializations.

4. EXPERIMENT RESULTS

4.1 Segmentation

Cell segmentation would be the first step toward automated analysis of histopathological slides. This is implemented in two steps in this study. In the first step, cell regions are segmented by clustering the pixel data Shown in Fig.2 and in the second step segmented cell regions are further processed by a watershed-based segmentation algorithm to identify individual cells. After the segmentation, feature extraction is done. At last perform classifier training and testing. Cell segmentation results in overlapped cells which further result in blobs, to separate these blobs we use watershed algorithm based on

immersion simulations. It is based on local maxima in the Euclidean distance mapped as seed points. These seed points are dilated either until the edge of the particle or the edge of the region of another growing seed point is reached.

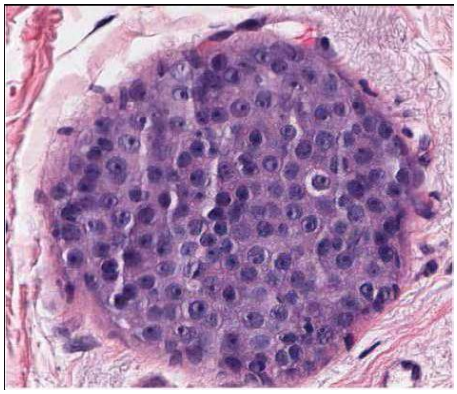

Fig2. (a) Test Image

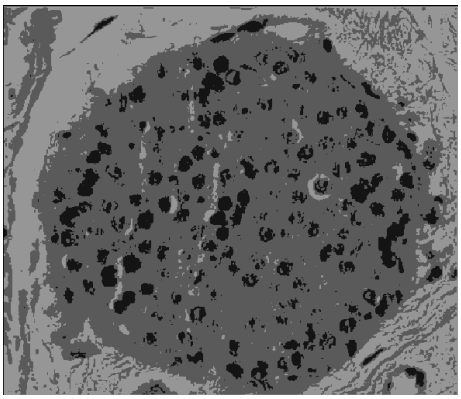

Fig.2(b) L-cell region segmentation

Once all catchment basins are identified and separated, the region defined by a catchment basin is considered a cell region.

4.1 Classification by GMM

GMM SVM method is based on the technique that each slide contains multiple ROIs and a positive (actionable) diagnosis is confirmed when at least one of the ROIs in the slide is identified as positive. For a negative diagnosis (UDH), the pathologist has to rule out the possibility of each and every ROI being actionable. The objective of this method is to develop a classifier to optimize classification accuracy at the slide level. For this supervised training techniques which are trained to optimize classifier performance at the instance level yield suboptimal performance satisfied by MIL.

5. CONCLUSION AND FUTURE WORK

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 detect 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] *M. M. Dundar, S. Badve, V. Raykar, R. K. Jain, O. Sertel, and M. N. Gurcan,* "A multiple instance learning approach toward optimal classification of pathology slides: A case study: Intraductal breast lesions," in Proc.20th Int. Conf. Pattern Recognit., Istanbul, Turkey, Aug. 23–26, 2010, pp. 2732–2735.

[8] *D.Wu, J. Bi, and K. Boyer,* "A min–max framework of cascaded classifier with multiple instance learning for computer aided diagnosis," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR 2009), Miami, FL, Jun. 20–25, pp. 1359–1366.

[9] *G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao,* "Multiple instance learning for computer aided diagnosis," in Advances in

Neural Information Processing Systems 19, B. Sch¨olkopf, J. Platt, and T. Hoffman, eds.Cambridge, MA: MIT Press, 2007, pp. 425–432.

[10] W. Rasband. (1997–2009). ImageJ, U.S. National Institutes of Health Bethesda, MD, Online:http://rsb.info.nih.gov/ij/

[11] T. G. Dietterich, R. H. Lathrop, and T. L. Perez, "Solving the multipleinstance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89,no. 1–2, pp. 31–71, 1997.

[12] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao, "Multiple instance learning for computer aided diagnosis," in *Advances in Neural InformationProcessing Systems 19*, B. Sch¨olkopf, J. Platt, and T. Hoffman, eds.Cambridge, MA: MIT Press, 2007, pp. 425–432.

[13] M. M. Dundar, S. Badve, V. Raykar, R. K. Jain, O. Sertel, and M. N. Gurcan, "A multiple instance learning approach toward optimal classification of pathology slides: A case study: Intraductal breast lesions," in *Proc.20th Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 23–26, 2010, pp. 2732–2735.

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

[15] www.ncbi.nlm.nih.gov

[16] www.medicinenet.com/breast_cancer/article.htm

[17] www.breastcancer.org/

[18] www.cancer.gov/cancertopics/types/breast

[19] www.cancer.org/cancer/breastcancer/index

[20] www.nationalbreastcancer.org

[21] www.breastcancercare.org.uk

[22] www.breastcancerindia.net

M.Sathish kumar is currently pursuing masters degree program in computer science and engineering in Karpagam University, Coimbatore,Tamilnadu, India. He received B.E in computer science and engineering from Angel College of Engineering and Technology, Tirupur. His area of interest includes medical image processing, networking and operating systems. He had presented more than eight papers in national level conferences and two papers in international conferences.