CANCER DETECTION USING HISTOPATHOLOGY IMAGES

Aiswarya P

Department of Electronics and Communication Engineering, Kuttippuram,
Aiswaryapmadhavan93@gmail.com

Dr.Mredhula L

Department of Electronics and Communication Engineering, Kuttippuram,
mredhu@yahoo.com

Abstract

In this paper we propose a method for the classification of histopathology images into Benign and malignant cancer. The system starts with the preprocessing of the image, which includes filtering and mask preparation of the images. The mask prepared in the preprocessing stage is used as the input of the segmentation process. Active contour without edge model is used for the segmentation. The features such as Area, Perimeter, Gray Level Co-occurrence Matrix (GLCM) are extracted. Linear Discriminate Analyzer (LDA) used as the classifier used and performance of which is compared with SVM and KNN classifier. The experimental results show that LDA classifier outperforms over others with 100% classification result. The compared result of the classifications helps for the identification of better classifier.

***Keywords*:** Biomedical histopathology images; active contour model; LDA; SVM; KNN.

Introduction

Breast cancer is one of the most common cancer among women in the worldwide. There are two states known as Benign, Malignant. Benign is slow growing and rarely spread to the other parts of the body, where is in the case of malignant it has faster growth on the body and is life threatening [1]. Due to the difficulty in understanding the breast cancer from the breast tumor most of the patients can't get suitable treatment at the best time. Diagnosis of the breast cancer mainly depends on the visual aspects of the tissue that are collected from the patients. In case of biopsy pathologist identify the mitosis by microscopic examination of stained slide in different magnification levels. Pathologist mainly focus on the size, shape, texture, spatial arrangements of nuclei and its interaction with stroma. These are more significant features that differentiate the benign and malignant cancer. Histopathology image is the image format of biopsy or surgical specimen. So, the analysis of the histopathology image is the process of doing cancer detection without human effort.

Histopathology images are prepared from the tissue slide. A relatively large proportion of the samples that are evaluated in pathology labs are from breast cancer patients. The study of immunohistochemically stained slides mainly includes the assessment of the number of cells that are positive for a specific antigen and the degree of positivity (staining intensity). The main objective of digitalization is to change the optical microscope as the principal tool used by pathologists. This process of implementation of digital slide images is somewhat equivalent to the digitization of radiological imaging. Previous approaches to histopathological image classification involve supervised learning techniques that use manually selected regions of interest with class labels provided by pathologists. However, these methods are not directly related to the analysis of new generation whole slide images that contain multiple areas with different levels of diagnostic importance. Thus, the identification and localization of diagnostically relevant regions of interest has appeared as an important initial step for whole slide image analysis. Different stages of the study include detection, segmentation, feature extraction, classification. In detection the identification of the location of each nucleus is take place and which is the leading stage of segmentation. If detection stage fails to identify all nucleus that there is degradation in result of segmentation.

There are different approaches in the literature to implement breast cancer detection. Tissue preparation and histopathology image formation are well stated by Mitko Veta and Josien in their paper [1]. Detection of breast cancer get poor result when the histopathology image is used without normalization and filtering. The difference in the result with and without filtering are explained in some of the papers [3]. The segmentation is the major and significant part of the breast cancer detection process. By using different segmentations, the accuracy of the classification varies. Watershed transform leads to over Segmentation of nuclei and which optimize the detection [4]. In detection process main part is segmentation, so selection of segmentation is more important in the section. In some of the cases different methods of segmentation are used and the result are combined for better result. In the combination process one method segments the nuclei from the image and the next segmentation process segment the stroma potion of the image [5]. From the analysis of most segmentation process the better result is for the active contour segmentation. Active contour segmentation mainly focuses on the minimization of energy in the contour. In some combines Active Contour Model (ACM) with Hough transform and Difference of Gaussian (DoG) filter are combined to detect cell nuclei. But it suits only for circular nuclei and require excessive computation [4]. In most of the current papers try to improve the histopathology image to get improved result in the detection process.

Coming to feature extraction it relates to which method will be used to extract features from the image character as representations. On the other hand, in feature selection, the most appropriate features to improve the classification accuracy must be searched. In the detection process of breast cancer, the main features used are texture and morphological features [6]. It is because the main difference in the benign and the malignant image is in their texture. Pathologist mainly focus on the shape, size, texture, alignment of nuclei in the stroma which can be identified from the area calculation of the nuclei. This paper mainly focuses on the detection of nuclei accurately, including mask preparation for the detection process.

1. Methodology

For the analysis and classification two types of cancer state images used are benign and malignant. The dataset of BreaKHis consist of images which are collected through a clinical study from January 2014 to December 2014. Samples are generated from breast tissue biopsy slides, stained with hematoxylin and eosin (HE). The samples are collected SOB, arranged for histological study, and labeled by pathologists of the P&D Lab. The preparation procedure used in this work is the standard paraffin process, which is usually used in clinical routine. Tissue is mounted on slide and sections of 3 µm are cut using a microtome. After staining, the sections are covered with a glass coverslip. An Olympus BX-50 system microscope with a relay lens of magnification of 3.3×coupled to a Samsung digital color camera.

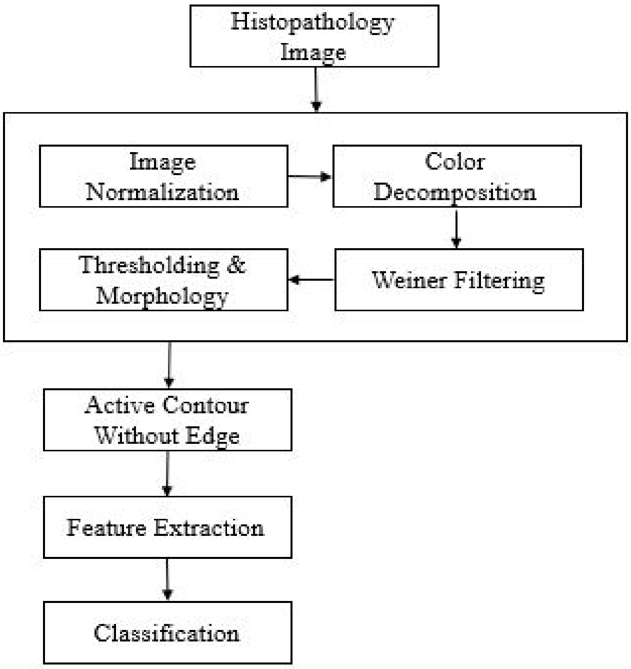

Fig .1 Block Schematic of proposed method

SCC-131AN is used to obtain digitized images from the breast tissue slides. The images are cropped and saved in three-channel RGB, 8-bit depth in each channel, portable network graphics format [7]. From the collection of data set, 35 benign and 35 malignant images are used for the analysis which have magnification of 200x and have resolution of 460 x 700 x 3.

The breast cancer detection includes mainly four steps. Preprocessing, Segmentation, Feature extraction, Classification. The process starts from the preprocessing stage. The preprocessing stage includes Normalization, Filtering and mask preparation. The mask points out the nuclei location from the image. The segmentation process starts from the preprocessed mask. From the segmented image the features for the classification is calculated. The feature extraction and feature selections are processed to get more better result in a short time. The classification includes two phases. The first phase is training, and the next phase is testing. The training stage compile the features for training the classifier. From the training data, the classifier identifies the major characters of the feature set of benign and malignant image. Next in the testing stage, which test the test images whether it is benign or malignant from the trained knowledge of the classifier. If the training image data increases, it increases the accuracy of the test process. It is because of the training stage, the classifier gets more information when increases the training dataset. The classification accuracy varies with varying the classifier.

1.1. *Stain Normalization and Mark Preparation*

Staining process have large variation due to the change in laboratories and change in the pathologist and the time of staining. So, dataset have different dynamic range and it required to normalize for further analysis. This avoid complications in quantitative assessment of histopathology images. Normalization is done by translating the color intensity value into linear form by converting the color value into optical density plane proposed by Marc Macenko et al. [6]. The normalized image is then decomposed into R, G, B plane. The R plane is only used for the next process. It is because the cell nuclei are darker in the R plane compared to G, B, plane. The R plane is then processed with wiener filter. The weiner filter is a linear filter and therefore it gives better filtering result compared to all other filters. Then the filtered image is used for the thresholding process which results in a binary image. The binary image is then proceeded to morphological operation. Accurate detection of nuclei is obtained by morphological reconstruction followed by thresholding. Morphological opening by reconstruction of the binary image is performed with a disc shaped structuring element. Disc shaped structuring element is used because the nuclei have disc shape. The reconstructed image expressed as:

$$\phi(I_B) = I_B + \rho_s(I_B \,|I_B)$$

Where $\rho_S(I_B \,|I_C)$ is the reconstructor operator with structuring Element.

1.2. *Active Contour Without Edge Segmentation*

The segmentation method is based on the minimization of energy. It is trying to identify the total energy of the curve and then change the curve until minimum total energyis get. Method considers the boundary of the object as the minimizer of the fitting term which is

$$F_1(C) + F_2(C) = \int_{inside\,(C)} \lambda_1 |I_0(x,y) - c_1|^2 dxdy +$$

$$\int_{inside\,(C)} \lambda_2 |I_0(x,y) - c_2|^2 dxdy$$

The case of energy calculation some time uses the area and length of the curve but here we have not used any area, length parameters.

The sum of internal and external energy terms is become minimum in the case of boundary of object. But if the curve is outside the object then the first term of the fitting term (Energy) become high and the second term is small. If the curve is inside the object, then the second term become high and the first term small. The analysis of the curve continues until energy term becomes as smaller as possible. The energy minimization identifies the all nuclei even if it is overlapped with each other. The minimum energy portions or the segmented parts of the image are the focusing part for further disease identification and classification. After detection of the nuclei we need to classify the detected cells whether benign or malignant. For this purpose, extract feature set from the segmented image.

1.3. *Feature Extraction*

In this work texture and morphological features are used for the classification of images. The main concentration in the feature extraction is to determine the difference in the nuclei of the benign and malignant cancer images. Area, Perimeter are useful for the determination of the difference in the characters benign and malignant nuclei. The other difference is the alignment of the nuclei with in the stroma. In the benign images the nuclei are aligned only around the stroma but in the case of malignant which are spread over all part of the tissue. By the analysis of this difference GLCM features can differentiate the difference of the benign and malignant images.

1.4. *Classification*

The classification stage the extracted features are trained and tested. The test result is evaluated using sensitivity, specificity, accuracy, precision, F-score. In the majority of breast cancer detection cases, complex classification methods are used. Linear discriminant classifier(LDC), K-NN (Nearest Neighbour) classifier and support vector machine classifier(SVM) are used in this work.

$$\text{Sensitivity} \qquad \gamma = \frac{TP}{TP+FN}$$

$$\text{Specificity} \qquad \delta = \frac{TN}{TN+FP}$$

$$\text{Accuracy} \qquad \varphi = \frac{TN+TP}{TN+FP+TP+FN}$$

$$\text{Precision} \quad \theta = \frac{TP}{TP+FP}$$

$$\text{Fscore} \quad = \frac{2x(\gamma x\theta)}{(\gamma+\theta)}$$

TP is defined as the number of true positives or correct detection of mitotic nuclei and FP as the number of false positives. Number of correctly detected non-mitotic nuclei is denoted as TN and number of wrong detection of non-mitotic nuclei are referred as FN.

2. Results

The result of the breast detection is compared by using different classifiers. Quantitative analysis of LDA classifier shows that it outperforms other methods and achieves 100% correct classification result. The SVM shows better result than the KNN classifier. SVM shows a classification accuracy of 85% and KNN shows 77% classification accuracy. The filtering and mask preparation are the performance improvement factors of the breast cancer detection. Results of the process are shown below (a)Original image (b)Normalized image (c)Segmented image.

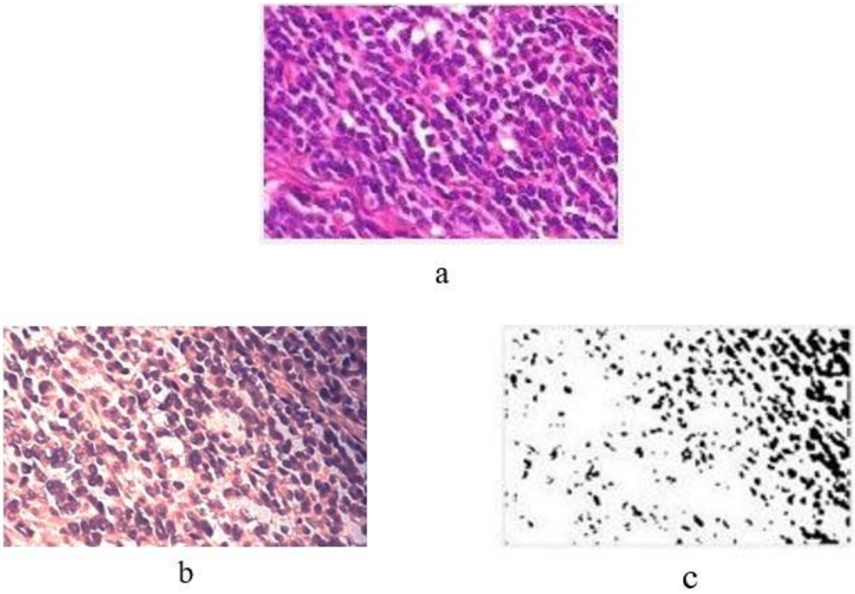

Fig. 2. (a) Original image. (b)Normalized image. (c)Segmented image.

The performance of the classification depends on the segmented output of the image. There is significant change in the normalized and original image, it increases the segmentation result.

Table 1. Classifier Performance.

Classifier	γ	δ	φ	θ	F-score
LDA	100	100	100	100	100
SVM	85	100	90.9	100	92
KNN	76.8	88	82	91	82.91

References

[1] Veta, M., Pluim, J. P., Van Diest, P. J., & Viergever, M. A. (2014). Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, *61*(5), 1400-1411.

[2] Beevi, K. S., Nair, M. S., & Bindu, G. R. (2016). Detection of mitotic nuclei in breast histopathology images using localized ACM and random kitchen sink based classifier. In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the (pp. 2435-2439). IEEE.

[3] F. T., & Shuvo, M. M. H. (2016). Detection of breast cancer from histopathology image and classifying benign and malignant state using fuzzy logic. In Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on (pp. 1-5). IEEE.

[4] Wang, P., Hu, X., Li, Y., Liu, Q., & Zhu, X. (2016). Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Processing, 122, 1-13.

[5] Chen, J. M., Qu, A. P., Wang, L. W., Yuan, J. P., Yang, F., Xiang, Q. M., ... & Li, Y. (2015). New breast cancer prognostic factors identified by computer-aided image analysis of HE stained histopathology images. Scientific reports, 5, 10690.

[6] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., ... & Thomas, N. E. (2009, June). A method for normalizing histology slides for quantitative analysis. In Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on (pp. 1107-1110). IEEE.

[7] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 63(7), 1455-1462.