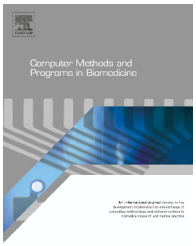
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 110 (2013) 298–307

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Performance comparison of machine learning methods for prognosis of hormone receptor status in breast cancer tissue samples

Adem Kalinli[a,*], *Fatih Sarikoc*[a], *Hulya Akgun*[b], *Figen Ozturk*[b]

[a] *Erciyes University, Engineering Faculty, Department of Computer Engineering, 38039 Kayseri, Turkey*
[b] *Erciyes University, Medical Faculty, Department of Pathology, 38039 Kayseri, Turkey*

ARTICLE INFO

Article history:
Received 15 May 2012
Received in revised form
25 December 2012
Accepted 26 December 2012

Keywords:
Image processing
Classification
Nucleus segmentation
Estrogen receptor (ER) status evaluation
Breast cancer prognosis

ABSTRACT

We examined the classification and prognostic scoring performances of several computer methods on different feature sets to obtain objective and reproducible analysis of estrogen receptor status in breast cancer tissue samples.

Radial basis function network, *k*-nearest neighborhood search, support vector machines, naive bayes, functional trees, and *k*-means clustering algorithm were applied to the test datasets. Several features were employed and the classification accuracies of each method for these features were examined. The assessment results of the methods on test images were also experimentally compared with those of two experts.

According to the results of our experimental work, a combination of functional trees and the naive bayes classifier gave the best prognostic scores indicating very good kappa agreement values ($\kappa = 0.899$ and $\kappa = 0.949$, $p < 0.001$) with the experts. This combination also gave the best dichotomization rate (96.3%) for assessment of estrogen receptor status. Wavelet color features provided better classification accuracy than Laws texture energy and co-occurrence matrix features.

1. Introduction

Breast cancer is the most common malignancy type among the female population, particularly in industrialized countries. Medical treatment of this disease is based on diagnostic and prognostic factors. One of the important prognostic factors used to define effective chemical substance for the patient, treatment duration and survival is evaluation of estrogen hormone receptor (ER) presence in the tumor section [1,2].

Immunohistochemical (IHC) staining of antibody-based assays is a commonly applied method to assess estrogen hormone receptor presence in tissue samples due to its relatively cheap material and easy procedure [3]. There are several scoring systems for the assessment of ER status [4–6]. In this work, ER status assessment was applied according to the Allred scoring system which is recommended as it is easy to use and is able to identify low-positive cases [7]. This scoring system calculates the proportion of stained cells, the staining intensity level of the sample and combines these results to give an overall scoring value [3].

Despite its advantages and widely availability in different laboratories, assessment of ER status is realized subjectively. Immunohistochemically stained breast tissue samples are

* *Corresponding author at*: Erciyes University, Scientific Research Projects Coordination Unit, 38039 Kayseri, Turkey. Tel.: +90 352 4379236; fax: +90 352 4379138.
E-mail addresses: kalinlia@erciyes.edu.tr (A. Kalinli), fsarikoc@erciyes.edu.tr (F. Sarikoc), hakgun@erciyes.edu.tr (H. Akgun), ozfigen@erciyes.edu.tr (F. Ozturk).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2012.12.005

visually examined by pathologists as colored images in a light microscope. Due to physio-psychological perception differences in human vision, an observer may give different score values for the same immunohistochemically stained image at different times (intra-observer variation) or different observers may give different score values for the same image at the same time (inter-observer variation). Extension of interpretation variation in medical practice and the need for a quality assurance scheme were addressed in a study by Rudiger et al. That study including 172 pathologists from Germany indicates that 24% of ER interpretations are false-negative [8].

Interpretation variation results not only from the human factor, but also from different laboratory conditions and IHC staining procedures. This diversity may be due to the fixation of tissue, duration and type of antigen retrieval, antibody specificity, antibody dilution and detection systems [6]. Therefore, IHC staining lacks standardization and reproducibility in medical laboratories [9].

In addition to the search for standard procedures in different laboratory conditions, several computer aided systems and methods have been presented in order to obtain more objective and reproducible results from IHC analyses. Motivated by this need, we conducted machine learning experiments on the same dataset by using different approaches in order to achieve better prognosis predictions. Additionally, we aimed to apply the Allred scoring protocol that has been recently adopted in Erciyes University Hospital.

Some of the previously reported methods in the literature make use of commercial software tools by generally implementing global thresholding techniques to obtain the luminosity in several color spaces [9–20]. In contrast to global threshold based methods, in the which stained pixel or area measurement of the image is considered, this study is based on detected nucleus perception and counting the nuclei as in real medical practice.

Other previously reported papers also relied on nucleus detection and benefited from texture features by employing k-nearest neighbors with weighted votes [21], radial basis neural networks [22], k-means clustering [23], probabilistic neural networks, support vector machines [24] and functional trees [25]. However, performance analysis of all these cited methods has not yet been reported using the same data set.

Here, the results of this experimental work revealed that a voting scheme combining functional trees and naive bayes has the best kappa agreement level with the assessments of observers ($\kappa = 0.899$ and $\kappa = 0.949$, $p < 0.001$) in ER status evaluation of test images. This scheme provided the best correlated scores with observer-2 ($\rho = 0.972$, $p < 0.001$), while the functional tree classifier gave the best correlated scores with observer-1 ($\rho = 0.963$, $p < 0.001$). Finally, the proposed scheme achieved a 96.3% correct dichotomization rate on the prognosis of test images.

In this paper, the methods and materials are described in Section 2. Results of feature selection, classification, assessments and statistical analysis are explained in Section 3. Discussion and conclusion are presented written in Section 4.

2. Methods

2.1. Materials

Invasive ductal carcinomas cases (40 specimens) from between 2007 and 2008 were identified from the files of the Department of Pathology at Erciyes University Medical Faculty in Turkey. Histological assessments were performed on 4–5 μm thick HE-stained sections of formalin-fixed paraffin-embedded tumors. Diaminobenzidine tetrahydrochloride (Dako Liquid DAB Plus, K3468, Denmark) was used as a chromogen and these sections were counterstained with Mayer's hematoxylin. Therefore, nuclei that have positive ER status expression were stained in brownish colors and nuclei with negative ER expression were stained in bluish colors. Staining of ER was evaluated in the nuclei of the malignant cells. The ER status was scored using the Allred scoring system [3].

Each slide was analyzed in a light microscope by the same pathologist (F.A.). She selected one representative region from each specimen under the microscope and captured the region in each slide with a linear magnification of 40 as a $2048 \times 1536 \times 24$ bit JPEG color image by means of a Lecia DMD 108 micro-imaging device. Considering inter-observer variation, two experienced pathologists, H.A. and M.K., manually assessed the images of these collected specimens according to the Allred scoring protocol. Those images with substantial visual artifacts, cytoplasmic stain or scoring disagreements of the experts (7 cases) were discarded because of usage limitation of the applied computer based assessment system [25].

From the remaining 33 cases, 6 cases were reserved as a separate training set and these were not used in test experiments. The experts were asked to mark some nuclei on the images of the training specimens by labeling each nucleus with a designated color by use of Microsoft Paint Brush. In total 384 nuclei were selected from these image files to from training data, when both of observers gave the same type of

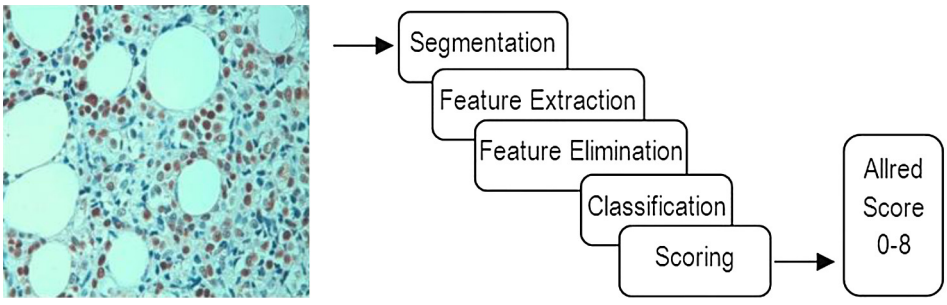


Fig. 1 – Processes of computer based prognosis system [25].