Breast Cancer Risk Estimation based on Machine Learning Methods for Computerized Assessment of Breast Composition in Digital Mammograms

Ya'akov Mandelbaum[1], Amitay Stein[2], Yitzhak Yitzhaky[2], Isaac Leichter[1]

1. Dept. of Applied Physics, Lev Academic Center, Jerusalem, Israel
2. Dept. of Electro-Optics Engineering, Ben Gurion University, Beer-Sheva, Israel

Objective:

The aim of the study is to develop a computer algorithm to automatically calculate the percentage of glandular tissue in a mammogram, making the results independent of the estimation of the interpreting radiologist.

Background:

A few studies have demonstrated a relationship between breast composition, tissue density in particular, and the risk of breast cancer [1]. Breast tissue which appears brighter on the mammogram is considered dense breast, and is due to a high percentage of glandular tissue. By contrast a high percentage of adipose (fatty) tissue in the breast reduces the breast density, and the resulting mammogram brightness. To date, the estimation of the percentage of glandular tissue is based on the subjective evaluation of the radiologist who must visually estimate the percentage of “bright areas" (glandular tissue) relative to the total breast image under consideration. This estimation is subjective and known to be imprecise and not consistent.

A typical mammography study contains four standard images, taken from different angles. In the MLO views the pectoral muscle, extraneous to the breast tissue, occupies a significant portion of the image. Any computerized analysis must start with the removal of the pectoral muscle from the image.

Material and Methods

The calculation of the percentage of glandular tissue was accomplished in two stages. First the subtraction of the pectoral muscle from the mammographic image was accomplished using a thresholding operation which creates a black and white

image in which the pectoral muscles appears differentiated from the adjacent breast tissue. The optimal threshold is determined by an algorithm which combines morphological methods with empirical results.

Following segmentation of the pectoral muscle, the glandular tissue is identified by classification of the mammographic images into 3 classes based on the characteristics of the histogram as well as texture analysis. For one class the glandular tissue was segmented using Seed Region Growing (SRG). For the other two classes, a threshold value was computed using a multivariate linear regression model, correlating histogram characteristics to an empirically specified threshold, determined by participating medical experts. Following identification of the glandular tissue, its area by percentage of the total breast tissue is computed.

Results:

The resulting algorithm was developed based on a training set, as described. Testing was performing on a verification set of 160 mammogram images. The results were compared to the area percentage computed based on the evaluation of independent radiologists, who manually defined the glandular tissue on the image. A high correlation of 0.92 was found between the results of the algorithm and those of the radiologists.

Conclusion:

The computerized algorithm developed presents an objective and systematic method to quantitatively evaluate the tissue density of breast tissue and thus improve the diagnostic accuracy of mammography.