● *Original Contribution*

COMPUTER-AIDED DIAGNOSIS FOR BREAST ULTRASOUND USING COMPUTERIZED BI-RADS FEATURES AND MACHINE LEARNING METHODS

JUAN SHAN,* S. KAISAR ALAM,[†‡§] BRIAN GARRA,[¶‖] YINGTAO ZHANG,[#] and TAHIRA AHMED[‖]

*Department of Computer Science, Seidenberg School of Computer Science and Information Systems, Pace University, New York, New York, USA; [†]Improlabs Pte Ltd, Valley Point, Singapore; [‡]Computational Biomedicine Imaging and Modeling Center (CBIM), Rutgers University, Piscataway, New Jersey, USA; [§]Department of Electrical & Electronic Engineering, Islamic University of Technology, Gazipur, Bangladesh; [¶]U.S. Food and Drug Administration, Silver Spring, Maryland, USA; [‖]Washington DC Veterans Affairs Medical Center, Washington, DC, USA; and [#]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Abstract—This work identifies effective computable features from the Breast Imaging Reporting and Data System (BI-RADS), to develop a computer-aided diagnosis (CAD) system for breast ultrasound. Computerized features corresponding to ultrasound BI-RADs categories were designed and tested using a database of 283 pathology-proven benign and malignant lesions. Features were selected based on classification performance using a "bottom-up" approach for different machine learning methods, including decision tree, artificial neural network, random forest and support vector machine. Using 10-fold cross-validation on the database of 283 cases, the highest area under the receiver operating characteristic (ROC) curve (AUC) was 0.84 from a support vector machine with 77.7% overall accuracy; the highest overall accuracy, 78.5%, was from a random forest with the AUC 0.83. Lesion margin and orientation were optimum features common to all of the different machine learning methods. These features can be used in CAD systems to help distinguish benign from worrisome lesions. (E-mail: jshan@pace.edu)   © 2015 World Federation for Ultrasound in Medicine & Biology.

Key Words: Breast cancer, Computer-aided diagnosis, Computerized features, Breast Imaging Reporting and Data System, BI-RADS, Machine learning, Receiver operating characteristic, Tissue characterization, Tumor classification, Ultrasonic imaging.

INTRODUCTION

Breast cancer is the leading cause of cancer-related death in women (Cheng et al. 2010). In 2014, approximately 232,670 new cases of breast cancer were diagnosed and resulted in approximately 40,000 deaths in the United States (Siegel et al. 2014). Screening mammography is widely used and recommended for the early detection of breast cancer. Studies have indicated that the addition of ultrasound can increase the overall cancer detection rate and reduce the number of unnecessary biopsies (Costantini et al. 2006; Hwang et al. 2005). Screening ultrasound is becoming an important addition to routine breast cancer screening because of its superior ability in imaging dense breast tissue and its lack of ionizing radiation.

Despite its many advantages, however, the quality of ultrasound has been relatively low because of the intrinsic speckle noise and low contrast between different tissue types. Digital image processing techniques and machine learning methods have been applied to improve detection rate and increase specificity (Chen et al. 2003; Huang et al. 2006; Segyeong et al. 2004). Advances in the field of medical image processing has improved the ability of computer-aided diagnosis (CAD) to reduce background noise, improve image contrast, detect regions of interest, differentiate a tumor from background and therefore help differentiate benign from worrisome lesions (Drukker et al. 2006; Moon et al. 2013a; Shen et al. 2007). Among all these functionalities of CAD systems, classifying a tumor into benign or worrisome categories is the ultimate objective.

The performance of machine learning methods relies heavily on how well the characteristics of tumors are represented by digital features, which can be separated into two categories: knowledge based and statistic based.

Address correspondence to: Juan Shan, Department of Computer Science, Seidenberg School of CSIS, Pace University, 163 William Street, New York, NY 10038, USA. E-mail: jshan@pace.edu

Knowledge-based features are derived from the Breast Imaging Reporting and Data System (BI-RADS) lexicon (Mendelson et al. 2013), which is used to characterize lesions based on shape, margin, orientation, echo pattern and acoustic shadowing (Chen et al. 2003; Moon et al. 2013a; Song et al. 2005). The other category of features is obtained from statistical computation, such as auto-covariance coefficients and frequency domain features (Huang and Chen 2005; Mogatadakala et al. 2006). These features capture the correlation between pixels and do not necessarily correspond to any observable features in ultrasound images.

The BI-RADS lexicon aims to standardize mammography and ultrasound reports so that reports are clear, succinct and consistent among readers. Although all BI-RADS terms are descriptive, not quantitative, they need to be "translated" into computerized features so a CAD system can compute these features automatically. Several groups have proposed approaches to quantify BI-RADS features (André et al. 2007; Mainiero et al. 2005; Moon et al. 2013b), including a comprehensive study by Alam et al. (2011). For example, the most commonly used ultrasound BI-RADS feature is the "parallel" orientation, which corresponds to the "long axis of a lesion paralleling the skin line." To quantify this feature, an equivalent ellipse of the lesion was identified, and the ratio between the horizontal axis and the vertical axis of the ellipse was computed (Chen et al. 2004; Moon et al. 2013a; Sahiner 2007). If the ratio is larger than one, the tumor is more likely benign; if the ratio is less than one, it is more likely malignant.

For this study, we performed a complete translation of the entire ultrasound BI-RADS lexicon into digital features, which are used in machine learning methods for the purpose of developing an effective CAD system for breast ultrasound. We have proposed new and validated digital features to distinguish benign from worrisome lesions with the ultimate goal of improving the accuracy of breast cancer diagnosis.

METHODS

The database used in this study contains 283 breast ultrasound images. The images were collected subsequently without excluding any data by the Second Affiliated Hospital of Harbin Medical University (Harbin, China), using a VIVID 7 (GE, Horten, Norway) with a 5- to 14-MHz linear probe. The aperture of the transducer is 4 cm. To obtain the original ultrasound images, the techniques harmonics, spatial compounding and speckle reduction were not used. The average size of the images is 500×420 pixels. The tumors range from 0.5 to 6.5 cm, with a median size is 1.1 cm. Among them, 133 cases are benign and 150 cases are malignant.

All lesions were validated by ultrasound-guided biopsy using a 14-gauge needle. Informed consent was obtained from all patients in this study. The study protocol was approved by the institutional ethics committee of the university. For each case, the boundary of the lesion was manually delineated by an experienced radiologist, and important findings were categorized into BI-RADS terms. The radiologist is a board-certified attending with more than 30 years of post-residency experience in breast ultrasound.

Computerization of BI-RADS features

In the fifth edition of BI-RADS Ultrasound (Mendelson et al. 2013), the dominant sonographic characteristics are grouped into five descriptive categories: shape, orientation, margin, echo pattern and posterior acoustic features. To quantify these BI-RADS features, multiple computerized features are proposed as discussed below and summarized in Table 1.

Shape. Irregular shape is a characteristic of malignancy. To capture this characteristic, an equivalent ellipse with the same second moments as the tumor region is used. As Figure 1 illustrates, the area difference between the tumor and its equivalent ellipse can describe how irregular the tumor is. This computable feature is called the area difference with equivalent ellipse (ADEE) and is defined as

$$\text{ADEE} = \frac{A_E + A^T - A_{E \cap T}}{A^T} \quad (1)$$

where A_E is the number of pixels in the equivalent ellipse, A_T is the number of pixels in the tumor region and $A_{E \cap T}$ is the number of pixels in the region where the tumor and the ellipse intersect.

Orientation. Orientation describes the direction of long axis of the tumor. If the long axis of the tumor parallels the skin line, the orientation is parallel, or "wider than tall"; otherwise, the orientation is anti-parallel, or "taller than wide." "Taller than wide" is a worrisome feature because malignant tumors have less compressibility and can grow across tissue planes. To quantify this feature, a minimum bounding box that covers the entire tumor is used. The edges of the bounding box should be parallel to the image boundaries. The ratio between the height and width of the box can characterize orientation as

$$R_{\text{Taller than wide}} = \frac{\text{height}}{\text{width}} \quad (2)$$

When the "taller than wide" ratio is larger than one, the tumor is likely to be malignant; otherwise, it is a sign of benignity.