Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images

Wen-Jia Kuo¹, Ruey-Feng Chang¹, Dar-Ren Chen², and Cheng Chun Lee³
¹Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi; ²Department of General Surgery; ³Department of Medical Research, China Medical College & Hospital Taichung, Taiwan

Key words: co-variance features, data mining, decision tree, ultrasound

Summary
To increase the ability of ultrasonographic (US) technology for the differential diagnosis of solid breast tumors, we describe a novel computer-aided diagnosis (CADx) system using data mining with decision tree for classification of breast tumor to increase the levels of diagnostic confidence and to provide the immediate second opinion for physicians. Cooperating with the texture information extracted from the region of interest (ROI) image, a decision tree model generated from the training data in a top-down, general-to-specific direction with 24 co-variance texture features is used to classify the tumors as benign or malignant. In the experiments, accuracy rates for a experienced physician and the proposed CADx are 86.67% (78/90) and 95.50% (86/90), respectively.

Introduction
The most useful way to reduce cancer deaths due to breast cancer is early detection and treatment. Thus, public awareness of the potential benefits of early detection of breast cancer has increased dramatically. Currently, the most frequently adopted methods for early detection of breast cancers are self-examination, mammography, and sonography. In spite of this there exist many kinds of detection methodology, biopsy is the best way to accurately determine whether the tumor is benign or malignant. More and more women are now seeking the advice of physicians immediately after detecting a breast mass. Consequently, surgeons perform an even increasing number of breast biopsies.

In comparison with other procedures, the disadvantage of biopsy is that it is more expensive for a large number of indeterminate lesions that intended to be differentiated per year. Therefore, it becomes constructive by diminishing invasive methods of distinguishing malignant from benign masses of breast to reduce the number of unnecessary biopsies, allay anxiety, and control costs. To reduce the unnecessary biopsies, mammography and sonography are supplied in most hospitals for doctors to diagnose via visual experiences. Biopsy is performed only when doctors suggested. Nevertheless, most biopsies are avoidable for the reason that the rate of positive findings at biopsy for cancer is very low, between 10% and 31% [1–3]. Furthermore, parenchyma deformity following excision biopsy may confound future mammograms and sonograms. Fine needle aspiration is the alternative way to overcome these shortcomings. In 1997, Rubin et al. [4] reported ‘fine needle aspiration biopsy should be the diagnostic procedure of choice for those patients classified clinically as probably benign or clinically as highly suspicious for cancer’. While ‘for patients classified as indeterminate, fine needle aspiration biopsy results are not reliable enough to determine treatment.’ Thus, how to increase the physician’s diagnostic confidence is the matter of concern. Recently, contrast-enhanced MR imaging [5, 6], breast scintigraphy [7, 8], and mammography [9–12] have all been discussed in recent reports as possible alternatives to excision biopsy for those benign and probably benign lesions. Another potential useful method of distinguishing benign from malignant masses of the breast is ultrasound [13]. Technical advances in US have
expanded the potential usefulness of this modality for the evaluation of breast lesions. Although nonpalpable and minimal tumors can be visualized by mammography, yet the US is a very convenient and safe diagnosis method. It performs very well, especially for the classification of palpable tumors. Recent works have emphasized the use of US for distinguishing benign and malignant solid nodules. In 1995, Stavros et al. showed that the sensitivity of breast ultrasound for malignancy is 98.4%, the specificity is 67.8%, and the overall accuracy is 72.9% [14]. The use of computer technology in decision support is now widespread and pervasive across a wide range of business and industry. Data mining becomes an important issue in many applications. This has resulted in the capture and availability of data in immense volume and proportion. With an acceptably accurate learning method, it is possible to develop predictive applications. Classification and regression are the main critical types of prediction problems. The classification solution is developed when the goal of prediction is discrete valued and the regression solution is developed when the goal of prediction is numerical or continuous. Predictive modeling methods have been developed drawing upon techniques from statistics, pattern recognition, and machine learning. This type of data analysis has been an active area of research in many scientific areas for a considerable period of time [15–18].

In this study, we evaluated breast masses in a series of pathologically proven tumors using data mining with decision tree model for classification of breast tumors. Region of interest (ROI) of US images and covariance texture parameters were used in our diagnosis system.

**Materials and methods**

There were 243 breast digital US images of pathologically proven tumors (either by cytology, core-needle biopsy or open biopsy) used in our study. Benign breast tumors from 161 patients (including 116 fibroadenomas, 25 fibrocystic nodules and 20 other benign lesions) and carcinomas from 82 patients (including 73 invasive duct carcinomas, five invasive lobular carcinomas, and four intraductal carcinomas). All these breast images were captured by one surgeon (D.R. Chen, who is also familiar with breast ultrasound interpretations) who also selected the ROI, data were consecutively collected from 1997 to 1998. Patients’ ages were ranged from 17 to 64-years old and tumor size measured from 0.8 to 4.2 cm.

**Data acquisition**

An ALOKA SDD 1200 scanner (Tokyo, Japan) and a 7.5 MHz linear real-time transducer with freeze-frame capability was performed to achieve the US images. No acoustic standoff pad was used with any of the cases. Digital US images were then obtained by the following steps. First, the analog signals from the VCR output of the scanner were transmitted to the frame grabber Video CATcher (Top Solution Technology; Taipei, Taiwan). Then, it digitized the data and generated frames of resolution 736 × 566 pixels for an NTSC (National Television Systems Committee) video-screen picture. Digital image was real-time captured and quantized into eight bits (i.e., 256 gray levels) using the Prolab's ProImage software package bundled with the frame grabber.

The physician (D.R. Chen) located ROI that contains the tumor from the full image and saved as a file with an impression before tissue proof was obtained for further analysis and diagnosis. Figure 1 illustrates an example of real-time digitized monochrome US image of a malignant tumor. ROI images are used in our breast image database to further investigate the texture characteristics of benign and malignant tumors.

**Image analysis**

Many researches in US texture analysis had been proposed over the last 20 years [19–26]. Texture analysis can be briefly classified in three main groups: the models, the mathematical morphology and the statist-

![736 × 566 digital ultrasonic image](image)

**Figure 1.** A 736 × 556 digital image is captured from the ultrasound scanner. In a 1 cm × 1 cm rectangle, there are 58 × 58 = 3364 pixels. The ROI rectangle is 2.28 cm × 1.88 cm and 132 × 109 pixels.
ical method. Markovian model [27, 28] and Fractal model [18, 29, 30] are the two most adopted models. Mathematical morphology methods are based on the study of shape representation in the image, yet these shapes are not always directly visible. Gray level morphology of images and morphological granulometry are the most two frequently used methods in mathematical morphology. The natural textures usually have a highly stochastic characterization whether they are structured or not. It would seem advisable to use statistical measurements in order to characterize those signals that are insufficiently described by most other approaches. Statistical parameters are evaluated either at order 1, at order 2 or at higher orders.

The most useful features were those directed from co-occurrence matrices of the image described as follows:

A co-occurrence matrix, $P_{\phi,d}(i,j)$, is a matrix where the $(i,j)$th element describes the frequency of occurrence of two pixels which are separated by distance $d$ in the direction $\phi$ with gray levels $i$ and $j$. For an $N \times N$ region with $M$ gray levels (0, ..., $M - 1$), the gray-level differences that single pairs of pixels can exhibit are:

$$(0, 0) \quad (0, 1) \quad \ldots \quad (0, M - 1)$$

$$(1, 0) \quad (1, 1) \quad \ldots \quad (1, M - 1)$$

$\vdots$ $\quad \vdots$ $\quad \vdots$ $\quad \vdots$

$$(M - 1, 0) \quad (M - 1, 1) \quad \ldots \quad (M - 1, M - 1)$$

We can find that the co-occurrence matrix can capture the texture variations in a region by various $\phi$ and $d$. The main power of co-occurrence matrix approach is that it characterizes the spatial interrelationships of the gray levels in a texture pattern and it is invariant under monotonic gray-level transformations. In general, a minimum set of co-occurrence matrices is four ($\phi = 0, 45, 90$ and $135$ degrees respectively; $d = 1$) for the texture that we have no prior knowledge about it. However, it is obvious that the size of the matrix is a function of the number of gray levels in the image and it would be prohibitively expensive to evaluate a matrix for each pixel in a general 8-bit image ($256 \times 256$ elements in a co-occurrence matrix). Thus, we use the statistical parameter matrices [26] with the same texture reservation properties instead of co-occurrence matrix.

The statistical method we adopted including contrast, co-variance, and dissimilarity that can evaluate the texture parameters for several distances between pixels and directly from the image without using co-occurrence matrices. The main advantage of this method resides in its calculation cost which depends only on the size of the image treated and not on the number of gray levels. Moreover, it allows the extraction of visually perceptible physical parameters from the image such as contrast, granularity, regularity, periodicity, fineness or coarseness of the texture, and so on. We briefly describe the evaluation of texture parameters as follows:

Let $S$ be a region of an image, $g(i, j)$ be the intensity value at the position $(i, j)$ of $S$, $\delta = (\Delta i, \Delta j)$ be the distance between two pixels and $\eta$ be the gray level average of the region $S$. For a value of $\delta$, contrast, co-variance and dissimilarity are evaluated in the following way:

**Contrast:**

$$\text{Con}(\delta) = E\{g(i, j) \cdot g(i + \Delta i, j + \Delta j)\}$$

**Co-variance:**

$$\text{Cov}(\delta) = E\{(g(i, j) - \eta) \cdot (g(i + \Delta i, j + \Delta j) - \eta)\}$$

**Dissimilarity:**

$$\text{Diss}(\delta) = E\{|g(i, j) - g(i + \Delta i, j + \Delta j)|\}.$$ 

According to our experimental results, we find that the co-variance performs better than other texture parameters. Thus, we adopt the co-variance as our texture features.

**Data mining with decision tree model**

Data mining has become a popular technology in the current researches for many applications. It is the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. The term ‘Data Mining’ refers to using a variety of techniques to identify numerous information or decision-making knowledge in bodies of data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, detecting anomalies, and so on. It extracts the information to be used in many areas such as decision support, prediction, forecasting, and estimation. The data is often voluminous with low value in its raw form that can be rarely made of it directly. However, the valuable part is the hidden information in the data.

In data mining, the greatest chance of success comes from combining expert’s knowledge of the data with advanced analysis techniques in which the computer itself identifies the underlying relationships and
features in the data. The process of data mining generates models from historical data are later used for prediction. With an appropriate learning method, it is possible to develop accurately predictive applications. The predictive modeling methods have been developed drawing upon techniques from statistics, pattern recognition, and machine learning. The techniques used to build these models are often referred to as machine learning or modeling. There are a number of machine learning and modeling technologies, including rule induction, neural networks, association rule discovery, clustering, and so on. In this paper, the decision tree model is used as the technique for mining the information we need for diagnosis.

A decision tree is a model that is both predictive and descriptive. The main reason that it is called a decision tree is the resulting model is presented in the form of a tree structure with decision rules. Decision trees are most commonly used for classification to predict what group a case belongs to. It can also be used for regression to predict a specific value. As a result, the decision tree has become a very popular data mining technique in many current applications. Thus, we chose the decision trees as the model in our study because the main purpose of our issue is to differentiate the tumors correctly.

The decision tree model encompasses a number of specific algorithms such as Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID), C4.5 [31] and C5.0 (from work by J. Ross Quinlan of Rulequest Research Pty Ltd, in St. Ives, Australia). In the AI field, C4.5 is one of the most popular inductive learning algorithms originally proposed by J. Quinlan. This algorithm constructs classification rules as a decision tree by some features and its relations that works based on the principle of expected information maximization. In our study, C5.0 algorithm which the speed and quality of rule generation are improved over its predecessor C4.5 is adopted to construct the decision tree model. In this paper, the co-variance texture parameters are used as the inputs to construct the decision tree model and for further diagnosis.

Results

In our experiments, 153 pathology-proven cases (including 101 benign breast tumors and 52 carcinomas) were sampled to be the training set for generating the decision tree model and 90 pathology-proven cases (30 carcinomas and 60 benign breast tumors) were sampled to evaluate the performance. Computer was used to analyze the sub-image extracted from ROI by a physician using the information about intensity variation and texture to make a differential diagnosis. Co-variance parameters were used as the feature of representing the texture properties of the ROI image to be diagnosed. The size of region for evaluating the co-variance parameters is $5 \times 5$ (25) pixels. C5.0 algorithm is used as the decision tree construction procedure. Furthermore, we compare the performance of the proposed CADx system with the physician (D.R. Chen).

Accuracy, sensitivity, specificity, positive predictive value and negative predictive value are the five most generally used objective indices to estimate the performance of diagnosis results. In our experiment, accuracy of data mining with decision tree for classifying malignancies was 96% (86/90), the sensitivity was 93.33% (28/30), the specificity was 96.67% (58/60), the positive predictive value was 93.33% (28/30), and negative predictive value was 96.67% (58/60) for the proposed CADx system. Table 1 lists the classification of breast nodules by the proposed data mining with decision tree model and an experienced physician. Sensitivity and specificity are the most two important indices that a doctor concerned. With sensitivity
Table 2. Summary of performance between an experienced physician and the proposed data mining with decision tree model diagnostic system

<table>
<thead>
<tr>
<th>Item</th>
<th>Physician</th>
<th>Data mining with decision tree model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>86.67</td>
<td>95.50</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>86.67</td>
<td>93.33</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>86.67</td>
<td>96.67</td>
</tr>
<tr>
<td>Positive predictive value (%)</td>
<td>76.47</td>
<td>93.33</td>
</tr>
<tr>
<td>Negative predictive value (%)</td>
<td>92.86</td>
<td>96.67</td>
</tr>
</tbody>
</table>

Accuracy = (TP + TN)/(TP + TN + FP + FN); Sensitivity = TP/(TP + FN); Specificity = TN/(TN + FP); Positive Predictive Value = TP/(TP + FP); Negative Predictive Value = TN/(TN + FN).

93.33% and specificity 96.67%, our proposed method provides objective evidences for good diagnoses of breast tumors. Summary of performance between an experienced physician and the proposed diagnostic system including the accuracy, sensitivity, specificity, positive predictive value, and the negative predictive value are listed in Table 2.

Discussion

Many lesions found at breast US are not clinically apparent by palpation, and in case of indeterminate lesions, this leaves the physician without diagnostic confidence and the patient with a high degree of anxiety. Improved imaging techniques have allowed the management of sonographic detected breast lesions to become less invasive. In this study, we proposed a novel diagnosis system in which inter-pixel correlation on the US images were used to differentiate benign and malignant tumors. Co-variance texture features and data mining with decision tree model were adopted to achieve better results. The potential of sonographic texture analysis to improve breast tumor diagnosis has already been demonstrated in the paper by Garra et al. [24] and by Goldberg [32]. A direct comparison between the methods and results of the present study with these earlier studies is therefore useful. This is a difficulty however, in that our study was not designed for this purpose. Difference in results between our study and previous ones could be due to either the specific population of cases examined or the different methods employed. Ideally, it would have been most useful if the methods of texture analysis used in the previous studies had been reproduced here, enabling a demonstration of whether the present data mining with decision tree model and autocorrelation function features do better than the other methods on the same set of data. If data from the selected ROI is still available, it may, even now, be possible to carry out such a comparative analysis. However it is inaccessible. Successful applications of the CAD strategy have been reported for other types of texture analysis [19–24], but differences in the evaluating results among this study and other ones is that we compared the performance between the ‘person’ and the ‘computer’. In this way, to claim its usefulness in clinical practice is trustworthy.

Clinically, physicians were not felt comfortable sometimes based on sonographic characteristic and thorough knowledge of clinical risk factors can lead to the categorization of patients into groups with relatively high rate of indeterminate or malignancy warranting biopsy. That causes the relative low positive predictive value (PPV) and in which cases only further clinical observation is appropriate. Applying mammographic technology and heightened sensitivity to breast disease have resulted in estimates of 500,000–1,000,000 breast biopsies performed yearly in the United States [33]. At the time of excision biopsy, 60–90% of lesions are benign [34]. Hall et al.[35] and Ciatto et al. [36] have attempted to increase the PPV of breast biopsies to a proposed goal of 41–51%. In order to forecast more precisely if a lesion is benign or malignant, breast ultrasound has been found to be helpful. In 1997, Raza used Doppler ultrasound to evaluate 86 solid breast lesions [37]. They concluded when gray-scale and power Doppler US findings were analyzed, 100% of the sensitivity and negative predictive value could be achieved while maintaining the 50% of PPV. Using this relatively simple image acquisition and computerized processing procedures, the PPV may be significantly improved from 76.40% to 93.33% and the discriminability of benign and malignant breast lesions is high. This CAD system can be readily adaptable to current US machine due to the simple use of standard output video and a video digitizer without extra effort or motion. Although we do not realize whether using the different ultrasonic machines can get the identical results, we believe that we can make it through the adjustment of the parameters by using intelligent selection algorithms according to the different rules achieved by retraining data obtained from different machines. This work is now ongoing. Owing to the higher PPV and NPV of this study, the number of biopsies of benign lesions by using these
methods can be lower down. This would increase the diagnostic confidence, offering a second reading to help reducing misdiagnosis and dramatically reduce the overall cost of diagnosing breast cancer in the practical circumstance. Thus far, this CAD system can enhance the diagnostic performance of radiologists, if it is used as a second opinion. However, CAD analysis has faced skepticism and numerous criticisms in the past. Its practical utility is controversial. Regardless, CAD is undergoing great development and utilization within the field of medical imaging by showing its potential in many clinical areas.

Acknowledgements

This work was supported by the National Science Council, Taiwan, Republic of China, under Grant NSC-89-2213-E-194-025.

References

31. Quinlan JR: C4.5 Programs for Machine Learning, Morgan Kaufmann, San Mateo, California, 1992

Address for offprints and correspondence: Dar-Ren Chen, Department of General Surgery, China Medical College & Hospital, 2 Yer-Der Rd, Taichung, Taiwan; Tel.: (886)-4-22052121 Ext. 1629; Fax: (886)-4-7267167; E-mail: dlchen88@ms13.hinet.net