Image Anal Stereol 2020;39:129-145 doi: 10.5566/ias.2113
Original Research Paper

STUDY OF CLASSIFICATION OF BREAST LESIONS USING TEXTURE GLCM FEATURES OBTAINED FROM THE RAW ULTRASOUND SIGNAL

MARIUSZ NIENIEWSKI✉,[1] AND LESZEK J CHMIELEWSKI[2]

[1]Faculty of Mathematics and Informatics, University of Lodz, ul. Banacha 22, 90-238 Lodz, Poland, [2]Institute of Information Technology, Warsaw University of Life Sciences - SGGW, ul. Nowoursynowska 159, 02-776 Warsaw, Poland
e-mail: mariusz.nieniewski@wmii.uni.lodz.pl, leszek_chmielewski@sggw.edu.pl

(Received January 27, 2019; revised January 2, 2020; accepted June 10, 2020)

ABSTRACT

Most of the methods of classification of breast lesions in ultrasound (US) images have been tested on B-mode images from the commercial equipment. The new possibility of further analysis of this problem showed up with the availability of a public database containing original raw radio frequency (RF) signals. In particular, it appeared that the original texture might contain diagnostic information which could be modified in the typical image processing and which is more difficult to perceive than the details of lesion shape/contour. In this paper a detailed analysis of the lesion texture is conducted by means of the decision trees and logistic regression. The decision trees turned out useful mainly for selecting texture features to be employed in the stepwise logistic regression. The RF signals database of 200 breast lesions was used for testing the performance of the benign vs malignant lesion classifier. The Gray Level Cooccurrence Matrix (GLCM) was calculated with the vertical/horizontal offset of up to five pixels. For each of these matrices six features were calculated resulting in a total of 210 features. Using these features a sufficient number of decision trees were generated to calculate pseudo-Receiver Operating Characteristics (ROCs). The outcome of this process is a collection of generated trees for which the employed features are known. These features were then used for generating generalized linear model by means of stepwise logistic regression. The analyzed regression models included the coefficients of up-to-the second degree terms. The texture features were further completed by a single shape feature, that is tumor circularity. The automatic procedure for finding the exact mask of a lesion is also provided for the conditions when the acoustic shadowing makes it impossible to obtain the entire contour reliably and a half-contour has to be used. The selected logistic regression models gave ROCs with the Area Under Curve (AUC) of up to 0.83 and the 95 % confidence region (0.63 0.96). Analyzing classification results one comes to the conclusion that the tumor circularity, which is the most informative among shape/contour features, is not essential against the background of textural features. The reported study shows that a relatively straightforward procedure can be employed to obtain benign vs malignant classifier comparable with that originally used for the database of the raw RF signals and based on the more complicated segmentation of the parameter maps of homodyned K distribution.

Keywords: breast lesion classification, quantitative ultrasound, feature selection, texture analysis, stepwise logistic regression.

INTRODUCTION

Breast cancer is the second leading cause of death for women all over the world. Up to recent times the most effective modality for detecting and diagnosing breast cancer has been mammography. However, mammography has its drawbacks, such as involvement of unnecessary biopsies due to false positive diagnosis, and health risk intensification both for patient and radiologist due to the use of radiation. The two problems that can be solved by using US images are: cancer detection and cancer classification. In the latter case the location and possibly the shape of the lesion have been obtained by some method and the task is to classify the lesion as benign or malignant. An extensive review of available methods for detection and classification of breast lesions in US images is presented by Cheng *et al.* (2010) and more recently by Menon *et al.* (2016). In the following mainly the most recent developments will be reviewed.

Harver *et al.* (2009) made use of features connected with lesion margin, that is margin brightness, margin sharpness, angular variation, as well as patient's age. Their classifier was based on logistic regression. Lee *et al.* (2009) described the shape of a lesion by means of a 1-D periodic signal and discrete periodized wavelet transform. Subsequently six morphometric features and 16 high octave energy features were calculated, and various combinations of the features were evaluated from the point of view of the efficacy of lesion classification.

Alvarenga *et al.* (2012) employed texture and morphological features for differentiation between benign and malignant lesions. Subsequently they applied the linear discriminant analysis to sets consisting of up to five features. Their results were obtained on a proprietary basis so direct comparison with other databases is impossible. Minavathi *et al.* (2012) performed detection, segmentation, and classification of lesions into spiculated and non-spiculated categories. Spiculations were detected by measuring the angle of curvature at each pixel of the mass boundary. The lesions were classified in accordance with measures defined for spiculations, elliptical approximation of the shape of the mass, and for acoustic shadowing caused by the mass. Zakeri *et al.* (2012) considered six features for classification of breast lesions: eccentricity, solidity, area difference of convex hull and rectangular box, area difference of mass and rectangular box, cross-correlation between rectangular box and its left-side (right-side) neighboring region.

Walach *et al.* (2013) based the segmentation of lesions on Maximally Stable Extreme Regions technique combined with posterior patches analysis. For each candidate from the segmentation step, a set of descriptors were calculated. The Support Vector Machine classifier was then used to distinguish between benign and malignant lesions. The method was tested on images collected from different acquisition devices. Moon *et al.* (2013) presented another approach which seems closer to what radiologists are doing. They segmented the images and calculated shape and texture features, 38 features altogether. Next they used the multinomial regression to obtain six BIRAD descriptive categories: shape, orientation, margins, lesion boundary, echo pattern, and posterior acoustic features. The quantified BIRAD findings were then used to determine the malignancy score for a lesion.

Nemat *et al.* (2018) investigated 21 shape-based features, 810 contour-based features and 24 texture features. They used the Bayesian extension of logistic regression with Automatic Relevance Detection Mechanism for rejection of irrelevant features. The obtained results were very good for a proprietary database.

The RF signals were also a subject of intensive investigation. The early paper by Lizzi *et al.* (1997) employed spectrum analysis of RF echo data to derive such features as attenuation, integrated backscatter, and sets of spectral parameters in order to assess tissue type. The authors analyzed probability density functions for each of these parameters for statistically homogeneous tissue structures such as prostate and liver. The papers by Alacam *et al.* (2003; 2004) developed the Fractional Differencing Autoregressive Moving Average (FARMA) model for RF signal. The FARMA model was used jointly with morphological features extracted from suspected areas for differentiating between benign and malignant lesions.

Granchi *et al.* (2015) proposed a hyper-spectral analysis method in which the spectral signal was decomposed into 16 sub-bands. The method used an automatic training procedure for which the coefficients of the decomposition were collected from selected areas in order to form clusters characterizing specific abnormalities. Having obtained the clusters one could classify any suspicious areas. The method was used for differentiating between infiltrating ductal carcinomas and fibroadenomas in breast tissue.

Uniyal *et al.* (2015) used ultrasound RF time series analysis as a method for classification of malignant breast lesions. Using the RF time series features and a machine learning framework the authors generated malignancy maps. These maps depicted the likelihood of malignancy for regions of size 1 mm within the suspicious lesions. The authors obtained AUC of 0.86 using Support Vector Machines and 0.81 using Random Forests classification algorithms. The frequency spectrum was estimated by calculating the FFT-based periodogram of the Hamming windowed time series. The estimated spectrum was divided into four frequency bands, each of which was averaged to deliver a feature. In addition to the RF time series features the authors also used texture features extracted from a B-mode image and spectral RF features extracted from an US frame.

Sadeghi-Naini *et al.* (2017) investigated the US spectral parametric maps in conjunction with texture analysis techniques to differentiate between benign and malignant breast lesions. The spectral analysis was performed on RF data and generated parametric maps of mid-band fit, spectral slope, spectral intercept, spacing among scatterers, average scatterer diameter, and average acoustic concentration. Subsequently the authors used texture analysis to determine mean, contrast, correlation, energy, and homogeneity features of the parametric maps. These features were utilized to classify benign vs malignant lesions with leave-one-patient-out cross-validation.

A separate group of publications deal with the open OASBUD database of raw RF US signals (Piotrzkowska-Wróblewska *et al.*, 2017) and include several papers by Byra *et al.* (2016) - Byra *et al.* (2018a) as well as an earlier paper by Piotrzkowska-Wróblewska *et al.* (2014). The advantage of this database is that it contains raw RF signals corresponding to benign and malignant lesions rather than their usual version obtained by log-compression and possibly some other auxiliary operations, for example, related to image resizing and hence implying interpolation, some filtering, etc. In particular, Byra *et al.* (2016) made use of statistical modeling of an US backscattered echo envelope for tissue characterization and developed the segmentation of homodyned K distribution parameter maps. Regions within lesions of different scattering properties were extracted and analyzed. Properties of these regions improved the distinction between benign and malignant tumors. In another paper, Byra (2018) considered two pattern recognition techniques. The first technique consisted in the US image eigen-decomposition and the use of the Fisher Linear Discriminant Analysis for differentiating between malignant and benign lesions. The second technique extracted the neural artistic style patterns of breast lesions using the VGG19 neural network. Next the Fisher Linear Discriminant Analysis was used to differentiate between style representations obtained for malignant and benign lesions. In another paper by Byra *et al.* (2017) the authors investigated how the shape and contour features of a lesion can improve the results of lesion classification when the features are used jointly with BIRADs assigned to a lesion by a radiologist. The classification was performed by logistic regression.

In a paper on CNNs. Byra *et al.* (2018a) described the transfer learning technique with the use of the InceptionV3 and the VGG19. Both neural networks were pre-trained on the ImageNet database and generated features used in the SVM classifier. The paper showed that the threshold chosen in the image reconstruction algorithm influences the results of the neural network with transfer learning and considered how to improve the classification.

The aim of the current paper is the investigation of classification of benign vs malignant breast lesions and evaluation of the discriminating value of the texture features calculated for the B-mode style images obtained from the raw RF signal not modified by any internal operations which may happen when one takes typical images obtained from the commercial US machines. Texture features have been successfully used in the analysis of medical images; for example, Pratiwi *et al.* (2015) worked with GLCM and Radial Basis Neural Networks for classification of mammograms. The additional objective of the current paper is to find out if there is any useful extra information in the shape/contour of the lesion in comparison with texture features. Two approaches to extracting information from the texture features are considered: stepwise logistic regression and decision trees. In fact the decision trees are not employed per se but rather as a tool for reduction of a number of features used by stepwise logistic regression. The supplementary purpose of the decision trees is to gain a feeling for what really determines the classification outcome in the situation when there are potentially hundreds of features, and we might end up with a complicated black box classifier, which nobody likes. In the context of the task of the current paper, it is interesting to note what is written on almost the same subject in the paper dealing with time series for the analysis of the B-mode images (Uniyal *et al.*, 2015). The authors of that paper state that the performance of B-mode texture features in breast cancer classification was poor in their study. They also say that the B-mode images used in their work were reconstructed from the RF signals offline. In contrast, the B-mode images from the commercial US machines are filtered and optimized in terms of dynamic range and have a higher quality compared to the B-mode images used by Uniyal *et al.* (2015). The authors' conclusion is that the reported performance of the B-mode texture features could potentially be improved by using the B-mode images produced by the scanner. The same authors also say that they performed a second kind of analysis where the entire lesion areas were considered as samples of unequal physical size for classification. And in this experiment, RF time series and B-mode and single frame RF features resulted in a higher AUC of 0.82 compared to B-mode and RF single frame features alone which resulted in AUC=0.68. In our view, this latter AUC seems rather low, and the current paper tends to improve on this. In fact, in our experiments we achieved AUC$\approx$ 0.83.

MATERIALS AND METHODS

DATABASE

An Open Access Series of Breast Ultrasonic Data (OASBUD) is a set of the raw RF ultrasonic echoes (RF signals) which were registered from 100 breast lesions (Piotrzkowska-Wróblewska *et al.*, 2017). In the set of 100 lesions, 52 solid lesions are malignant and 48 are benign. In the group of malignant lesions all lesions were histologically assessed by core needle biopsy. In the group of benign lesions part of them were histologically assessed, and part were observed over a two-year period.

A more detailed information supplied by the developers of the OASBUD database is the following. For each lesion, two individual orthogonal scans, called, respectively, Rf1 and Rf2, from the pathological region were acquired using an Ultrasonix

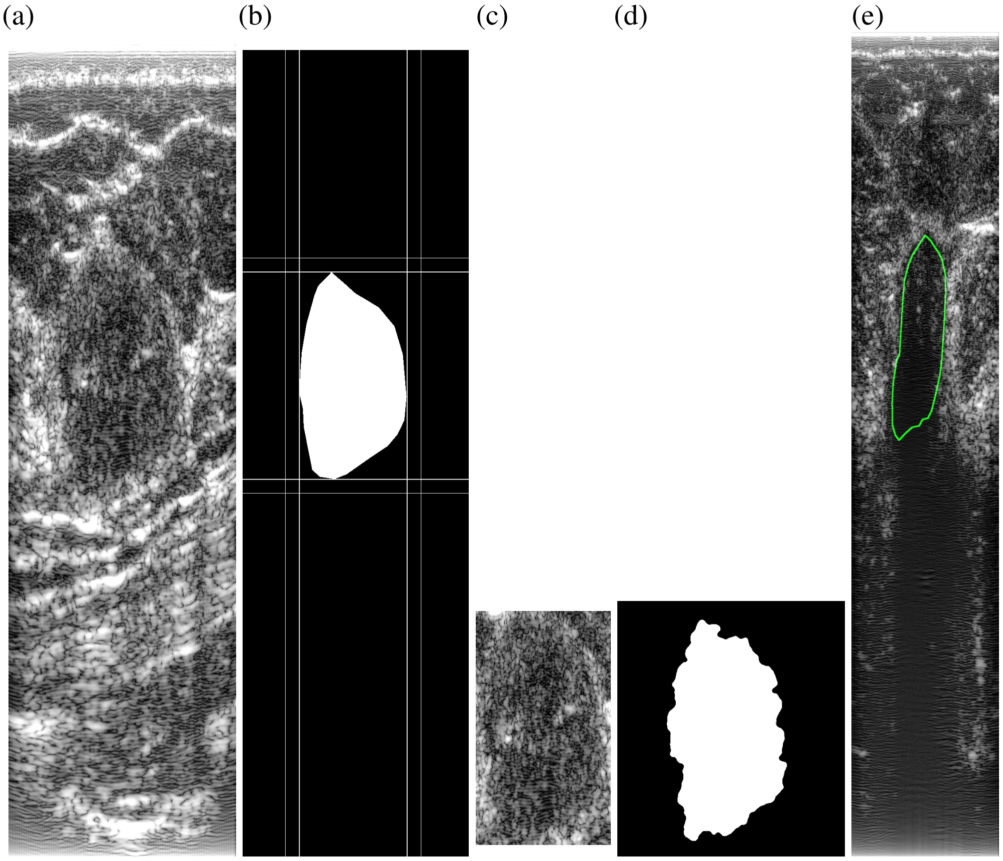

Fig. 1: Two examples of images from the OASBUD database. (a) Original image with a well-defined boundary. (b) Binary mask of the lesion from (a) obtained by a radiologist together with lines marking the position of the processing windows. (c) Smaller window used for the extraction of texture features. (d) Larger window used for the extraction of shape/contour features. (e) Original image with posterior acoustic shadowing. (Green line shows the contour of the mask drawn by a radiologist.)

SonixTouch Research ultrasound scanner with an L15-4 linear array transducer of center frequency 10 MHz. It is worth noting that a usual inexpensive clinical scanner would not give access to the RF signal. The focusing region of each scan was always placed at a depth of a lesion. Each image was reconstructed using 510 RF backscatter echoes lines. Signals were digitalized with 40 MHz sampling frequency. The number of samples in every RF signal depended on the chosen penetration depth. For example, for 40 MHz sampling rate the distance between adjacent samples was 0.0192 mm (for assumed sound speed in tissue 1540 m/s). As a result there would be 2596 samples for 5 cm penetration depth and 1558 samples for 3 cm depth. Other settings used by the operator had no influence on character of the raw ultrasonic echoes.

We obtained the B-mode type images by taking the Hilbert transform of the original RF signals. More details on this subject are given below in the Appendix: Essential Matlab Commands. The images have brightness in the range [0, 255]. However, the relevant information is contained in the range [40, 80] so that this range had its dynamics extended linearly to the range [0, 255]. The resulting images were not resized in any way so there was no interpolation involved. For our purposes the database is treated as more or less uniform collection of 200 RF signal samples, or images, representing 104 malignant and 96 benign cases. Examples of images and windows generated for the OASBUD database are shown in Figs. 1 and 8. In principle the window for calculating texture features should include the whole of a lesion, but it should be as small as possible in order to maintain the discriminatory power of the texture features, and it is called a smaller window here. In our calculations we added 20 pixels on each side of the bounding box of the mask, as illustrated by auxiliary lines in Fig. 1(b) and also by 1(c). Furthermore, we assumed another slightly larger window for calculating shape features of a lesion. The reason for the window's enlargment is that the database contains a rough polygonal approximation of the lesion contour that is not precise enough to be used for extraction of shape features. Obtaining a more precise contour may result in a bigger mask of the lesion. In order to allow for this increase it was necessary to enlarge the window. We found out that adding 20 pixels at the top and 20 pixels at the bottom of the smaller window was adequate for the whole database. The width of the larger windows was assumed for simplicity equal to the full width of the original image in the database.

TEXTURE FEATURES

In the current paper the texture features are obtained following definitions given in the book by Gonzalez *et al.* (2009). The GLCM specifies how many pairs of pixels in the image have a given doublet of gray levels. Any combination of gray levels may be assumed, and we use gray levels in the range 0 through 255, which gives us matrices of size 256×256. The pairs under consideration are chosen based on the vertical and horizontal offset between the pixels in the pair. Below, a number of different offsets are considered, and for each offset there is an individual GLCM. We assume that the calculations do not make difference depending on which pixel has which brightness; in other words, the pair of brightness values $(3,5)$ and $(5,3)$ are treated the same.

For a given offset and its corresponding GLCM, the features are calculated in the following order: **contrast, correlation, energy, homogeneity, maximum probability,** and **entropy.** The first four features are calculated as described in the book by Gonzalez *et al.* (2009). The last two features are calculated for the normalized GLCM as described in the same book (comp. Appendix: Essential Matlab Commands).

We assume 35 possible offsets, and for each offset there are six above mentioned features. The offsets are ordered as shown in Table 1. It means that we have $6 \times 35 = 210$ texture features. Having a particular feature index (#) it is easy to identify its name and the corresponding offset. For example for feature # 181 we are in the 31-st square in Table 1. That means we have offset (5, 4) and the first feature in the square is contrast.

Table 1: Texture feature ordering as a function of the offset in the GLCM. The three numbers in each square denote, respectively, vertical offset, horizontal offset, and the index of a 6-tuple of features.

5, 0 35	5, 1 34	5, 2 33	5, 3 32	5, 4 31	5, 5 30
4, 0 24	4, 1 23	4, 2 22	4, 3 21	4, 4 20	4, 5 29
3, 0 15	3, 1 14	3, 2 13	3, 3 12	3, 4 19	3, 5 28
2, 0 8	2, 1 7	2, 2 6	2, 3 11	2, 4 18	2, 5 27
1, 0 3	1, 1 2	1, 2 5	1, 3 10	1, 4 17	1, 5 26
0, 0	0, 1 1	0, 2 4	0, 3 9	0, 4 16	0, 5 25

SHAPE FEATURES

There are many possible shape/contour features. Below we concentrate on the tumor circularity whose value depends on whether the lesion is malignant or benign (Zhou *et al.*, 2015). However, the problem is complicated by the occurrence of the posterior acoustic shadowing, which may be observed for many malignant lesions. Piotrzkowska-Wróblewska *et al.* (2017) affirm that the RF signals in the OASBUD database were acquired in the way to minimize the amount of shadowing and artifacts affecting the quality of recorded data. However, for large lesions exceeding 40 mm, it was not always possible to expose clearly the lower edge of the lesion. Piotrzkowska-Wróblewska *et al.* (2017) sidestepped the problem by delivering the masks for all of the lesions. However, these masks were drawn by a radiologist and contain important information that is hard to extract automatically from the image.

The problem of acoustic shadowing is not new in the literature. For example, an early paper by Drukker *et al.* (2003) considered the influence of the acoustic shadowing on the detection of lesions. A more recent paper by Padilla *et al.* (2013) contains several examples of posterior acoustic shadowing in the case of benign breast lesions. An instructive collection of US lesion images with acoustic shadowing was also published by Gokhale (2009). It is obvious that the acoustic shadowing complicates classification of lesions since we cannot assume that we will have a reliable contour in every case.

The following considerations regarding acoustic shadowing were inspired by the work of Zhou *et al.* (2015) who made a point that the breast lesion segmentation and classification of US images may be biased by shadowing. As a remedy, Zhou *et al.* (2015) proposed the use of half-contour features for classification of lesions rather than a full contour. In particular, they analyzed the following six features: tumor circularity, mean of the normalized radial length, standard deviation of the normalized radial length, area ratio, roughness index, and standard deviation of degree. Zhou *et al.* (2015) showed that among these six features, the tumor circularity and standard deviation of degree for half-contour were most effective for classifying lesions with or without posterior acoustic shadowing. Following these findings we started with the tumor circularity (denoted TC below), which is a gross contour feature descriptor, calculated using the equation

$$\mathrm{TC} = \frac{P^2}{A}, \tag{1}$$

where P is the perimeter and A is the area of the lesion half-mask. The perimeter P is measured by summing the pixels belonging to the tumor contour, and the area A is calculated as a number of pixels inside the contour. Zhou *et al.* (2015) obtained the half-contour by determining the leftmost and rightmost pixels of the lesion region. Then a line was drawn connecting the above mentioned extreme pixels and the half-contour was defined as the upper portion of a lesion, above the line connecting the leftmost and rightmost pixels. We took a similar approach adding some specific steps described in detail in the Appendix: Obtaining the Upper Half-Mask below. The aim of these steps is to make the whole process of half-contour generation fully automatic and void of a manual intervention which otherwise would be necessary in case of some images.

RECEIVER OPERATING CHARACTERISTICS

There is a vast literature on ROCs. In this paper we follow the definitions formulated by Adler and Lausen (2009) who explained the idea in a very systematic manner. The necessary bootstrap approach was fully considered in a book by Efron and Tibshirani (1998). The early application of the bootstrap to US image analysis was given in a paper by Chen *et al.* (2002). Below only some basic ideas are gathered. A sample Z is given consisting of N observations $(x_i, y_i), i = 1, \ldots, N$. These observations are realizations of the random variables X and Y, where X is a l-dimensional vector of predictors, or features, and Y represents the class membership of individual feature vectors. In classification we want to predict the class for a given vector of predictors X. In a 2-class problem $Y \in \{0, 1\}$. The performance of the classifier is assessed in terms of the true (TPR) and false (FPR) positive rates, where TPR is a proportion of positives that were classified correctly, and FPR is a proportion of negatives that were classified as positive.

The true positive rate $\mathrm{TPR}(Th)$ is defined by Adler and Lausen (2009) as

$$\mathrm{TPR}(Th) = P[P(Y=1|X) \geq Th|Y=1], \tag{2}$$

where Th is some threshold. In this equation the internal $P(Y|X)$ is some classification rule $\hat{f}(x_i)$ and gives the estimate of the probability that X belongs to class 1. The false positive rate $\mathrm{FPR}(Th)$ is defined by a similar equation

$$\mathrm{FPR}(Th) = P[P(Y=1|X) \geq Th|Y=0] \tag{3}$$

The ROC is then obtained as

$$\mathrm{ROC}(.) = \{(\mathrm{FPR}(Th), \mathrm{TPR}(Th)), \quad Th \in [0\ 1]\} \tag{4}$$

where the limits for the threshold can be modified if necessary.

There are several varieties of the true and false positive error rates, hence several definitions of the ROC. In the following we will consider three such varieties (Adler and Lausen, 2009):

- apparent $\mathrm{TPR}(Th)$ and $\mathrm{FPR}(Th)$,
- bootstrap estimated $\mathrm{TPR}(Th)$ and $\mathrm{FPR}(Th)$,
- 0.632 bootstrap estimated $\mathrm{TPR}(Th)$ and $\mathrm{FPR}(Th)$.

For each pair of $\mathrm{TPR}(Th)$, $\mathrm{FPR}(Th)$ there is a corresponding ROC curve, termed respectively apparent, bootstrap estimated, and 0.632 bootstrap estimated ROC. Theoretical considerations and statistical experiments, for example in the paper by Adler and Lausen (2009), show that the most precise of these ROCs is the 0.632 bootstrap estimated ROC. In the following, examples of all three ROC curves, renamed as pseudo-ROCs, are shown in Fig. 2.

DECISION TREES

The decision trees are described in many publications. Here we mention only a review paper (Kotsiantis, 2013) and a book (Hastie *et al.*, 2009). In the case of the decision trees there is no single parameter that could serve for thresholding, and this threshold is the essence of ROC constructing since by reducing the threshold we are moving up on the ROC curve. However, a "pseudo-ROC" can be obtained by changing appropriate element in the cost matrix used for generating the tree. This cost matrix M has the form

$$M = \begin{bmatrix} 0 & 1 \\ acost & 0 \end{bmatrix}. \tag{5}$$

In the equation above, **zero** is the cost for the correct classification; **one** is the cost for classifying the benign case as malignant; *acost* is the cost for classifying malignant case as benign. In the particular case of $acost = 1$ the costs of erroneous classification for benign and malignant cases are equal.

The main idea of our approach was as follows. A number of trees were obtained by changing the minimum allowable number of observations per leaf, denoted here $MinLeaf$. We assumed this number to be in the range from 1 through 30. Then for each tree the *acost* in the cost matrix M was changed in the range from 1 through 35. A larger $MinLeaf$ results in a smaller tree. Then by increasing the cost of misclassifying the malignant case we increased the number of TP cases and hence moved upwards along the ROC generated for a given $MinLeaf$.

We split 200 (smaller) windows from the OASBUD database into two sets: training set containing 80 benign cases (class **zero**) and 80 malignant cases (class **one**), as well as testing set of 40 cases containing 16 benign and 24 malignant cases. This testing set was not used in connection with decision trees but was essential at later stage for logistic regression.

The apparent pseudo-ROCs were obtained by changing $acost$, and there was a separate ROC for each value of $MinLeaf$. The same data, that is 160 vectors of 210 features, were used both for tree generation and for classifier testing. This is in accordance with the definition of the apparent ROC.

Computations for bootstrap estimated pseudo-ROCs were significantly more complicated. We obtained 200 bootstrap samples each of 160 feature vectors by sampling with replacement from the original population of 160 feature vectors mentioned above. Each of these bootstrap samples served as a training set. In contrast, each respective testing set was obtained as a collection of all the feature vectors taken from the population of all 200 vectors that were not included in the given training set. Next we assumed a number of $MinLeaf$ values as described above. For each of these, a pseudo-ROC was calculated. The calculations ran as follows. In order to get a single point on the pseudo-ROC we selected a value of *acost* from a given range. For this *acost* we calculated the TPR and FPR for each bootstrap sample. The mean values of TPR and FPR over all bootstrap samples were used as coordinates of a single point on the pseudo-ROC. Repeating calculations for all values of *acost* resulted in one pseudo-ROC curve.

Computations for 0.632 bootstrap estimated pseudo-ROCs followed description given by Adler and Lausen (2009). These computations are quite straightforward and are not detailed here. However, it is worth noting that the 0.632 bootstrap pseudo-ROC takes into account the fact that in a real-life situation some number of the feature vectors used for teaching may be used for testing as well.

LOGISTIC REGRESSION

Logistic regression is profusely described in several books (Hastie *et al.*, 2009; Kleinbaum and Klein, 2010; Bingham and Fry, 2010; McCullagh and Nelder, 1989), and we do not enter into details here.

The logistic regression calculations were executed in three steps. The first step was generation of a generalized model using stepwise logistic regression that determined which features should be included and which should be removed from the model. Generation of the model requires the input in the form of the feature vectors and the desired labels for these vectors. The label is **zero** or **one** depending on whether a given vector represents a benign or malignant case. The features to be processed by the stepwise regression were pre-selected based on the information obtained from the decision trees. In fact there is certain flexibility in selecting the features and the best set(s) of features were chosen by observation of the final outcome of the classifier, which means looking for the highest AUC.

The second step was the calculation of predicted class labels for the above mentioned 40 testing set feature vectors including 16 benign cases and 24 malignant cases. These testing vectors were put aside when generating the decision trees so they did not influence selection of features for logistic regression in any way.

The third step consisted in actual calculation of the pseudo-ROCs using the results from the previous steps.

The actual calculations of all three steps of logistic regression are illustrated in the Appendix: Essential Matlab Commands.

RESULTS

RESULTS FOR DECISION TREES

The pseudo-ROC curves for exemplary values of $MinLeaf = 10, 18,$ and 30 are illustrated by the three left-most diagrams in Fig. 2. The last diagram in Fig. 2 is of a different kind. It shows the 0.632 bootstrap estimated ROC curves corresponding to red lines in the three left-most diagrams, but with the inclusion of the

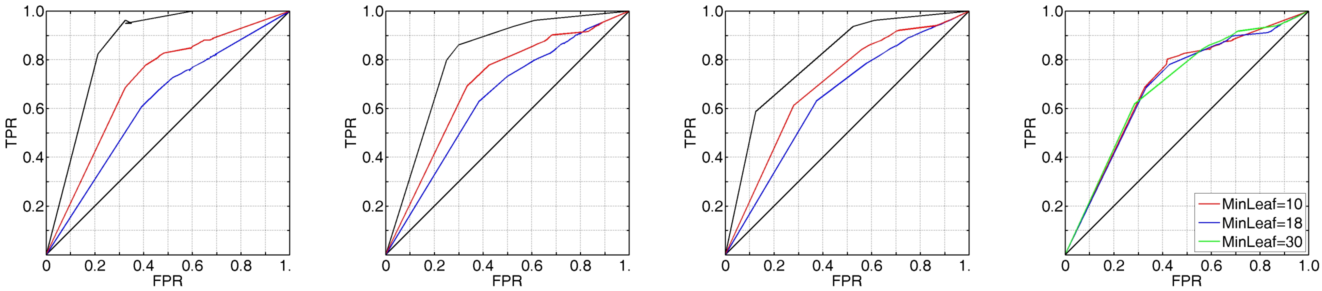

Fig. 2: Pseudo-ROCs obtained for the decisions trees. Three left-most diagrams represent pseudo-ROCs for texture features only. Black line is the apparent pseudo-ROC; blue line is the bootstrap estimated pseudo-ROC; red line is the 0.632 bootstrap estimated pseudo-ROC. The fourth, right-most diagram shows 0.632 bootstrap estimated pseudo-ROCs for texture features together with tumor circularity.

tumor circularity. Comparing red lines from the three left most diagrams to the lines in the last diagram one concludes that the influence of the tumor circularity on the shape of the ROC is minimal. The decision trees from Fig. 2 can be used for classification of breast lesions. Nevertheless, their quality is rather modest in the sense that they are distant from the upper left corner of coordinates (0, 1) of the square visible in all diagrams in Fig. 2. In the next section we will show how the information obtained from the tree generation, that is a set of features used by a tree, can be used in developing a stepwise logistic regression model, which gives an ROC superior to the pseudo-ROCs obtained for the decision trees.

As an example, specifications of the features selected by the trees generated for $Minleaf = 10$ and 18 and for $acost = 1$, for which an apparent ROC can easily be obtained, are given in Table 2.

Table 2: Examples of the texture features used in the generated trees.

Feature	Vertical offset	Horizontal offset	Feature #
$Minleaf = 10, acost = 1$			
Contrast	5	4	181
Homogeneity	4	1	136
Homogeneity	3	0	88
Contrast	4	4	115
Contrast	0	1	1
$Minleaf = 18, acost = 1$			
Contrast	5	4	181
Homogeneity	4	1	136
Homogeneity	3	0	88

A complete set of all the features used in decision trees obtained for $MinLeaf$ = 10, 18, and 30 is shown in Table 3.

RESULTS FOR LOGISTIC REGRESSION

The set of $MinLeaf = 10, 18, 30$ values in Table 3 roughly covers the whole range of interesting ROCs. 12 examples of the generalized linear regression models generated by stepwise regression are collected in Table 4. Each row in this table represents an individual model. The Features column specifies the features used by the given model. Feature # 211 is the tumor circularity. The odd rows represent models with no tumor circularity. The subsequent even rows represent, respectively, the same models as odd rows, but with the addition of the tumor circularity. All the models were generated by allowing an intercept, linear terms, interactions, and second order terms (mixed or squared). The starting model was of the first order. The Regression Model column in Table 4 shows the generated model in the shorthand notation used by Matlab, in which the actual regression coefficients are omitted. The optimal operating point (OP) is also known as a Younden index in the statistical literature. The AUC column in Table 4 specifies the Area Under Curve, with the first entry indicating the mean value, the second entry denoting the lower bound of the 95 % confidence interval, and the last entry denoting the upper bound of the same interval.

As an example, the actual regression equation for the model defined in the first row in Table 4 is as follows

$$y = 58.615 - 0.041377X_{181} - 273.05X_{136} \\ +165.89X_{88} + 8.1615 \times 10^{-06}X_{181}^2 \qquad (6)$$

The calculated y value is a random variable defined as a function of X variables. This value is subsequently mapped into probability $\mu = exp(y)/(1 - exp(y))$ used as a parameter for the binomial distribution, the samples from which are the generated labels.

Further parameters obtained for the regression

Table 3: All features obtained for three selected values of *MinLeaf*.

MinLeaf	*acost*	Feature #	Remarks
10	1	181, 136, 88, 115, 1	
10	2	121, 22, 136, 67, 88, 42, 33	
10	3	121, 22, 136, 67, 82, 88, 33	
10	4-5	121, 40, 133, 1, 82, 115, 145	
10	6-35	121, 40, 133, 1	
18	1	181, 136, 88	
18	2	121, 136, 88	
18	3	121, 136	
18	4-10	121, 133	
18	11-35		FPR=TPR=1
30	1	181	
30	2-3	121, 136	
30	4-10	121, 133	
30	11-35		FPR=TPR=1

model # 1 from Table 4 are given in Table 5, which specifies the Standard Error (SE), tStat, and pValue for all the coefficients in four exemplary regression models. It is typically assumed that the pValue should be less than 0.05 for a given estimate to be reliable. Inspection of Table 5 reveals that this condition is satisfied for almost all the coefficients, with two exceptions, of the models 5 and 7, in which the pValue for one of the coefficients is approx. 0.067. This indicates that these coefficients are on the verge of statistical significance. The only way of improving the pValues would be to use more images in the teaching set, but this would reduce the number of images available for the testing set. The models 1 and 11 have all pValues below 0.05 level. In summary, the models specified in Table 5 are close to the boundary of their statistical significance. The tStat values are not discussed here, but they can be calculated from pValues based on the fact that they both refer to the same t distribution.

Some examples of ROC curves generated for regression models from Table 4 are shown in Figs. 3-6. Fig. 3 corresponds to a single ROC. For visibility reasons the ROC curve itself is shown on the left, and the confidence intervals are added on the right. All other figures represent pairs of ROC curves together with their 95% confidence intervals. The optimal operating point is marked by a red circle. However, for comparison purposes more important is the AUC under the ROC, which is a single number characterizing the whole ROC curve. The values of AUCs are specified in Table 4 as well as in the captions of Figs. 3 - 6. The best AUC seems to be 0.83 with 95 % confidence interval (0.63 0.91), as shown in Fig. 5. However, the AUCs for other figures are very similar. The conclusion is that changing the regression model does not change the `AUC` significantly.

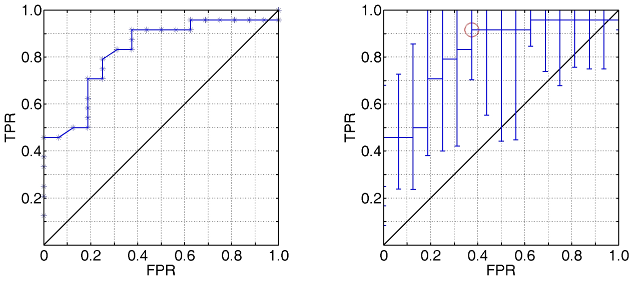

Fig. 3: Left: ROC curve for the logistic regression model # 1 in Table 4. Right: the same ROC curve with 95 % confidence intervals calculated by means of bootstrap sampling. AUC≈ 0.82. The 95 % confidence interval (0.64 0.93).

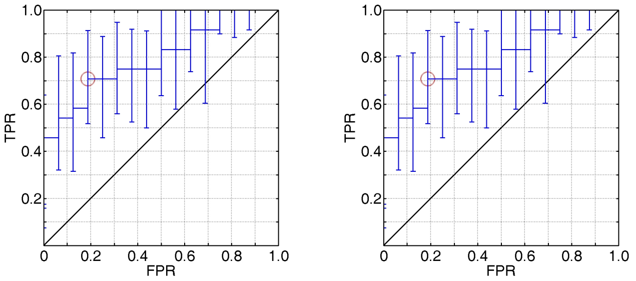

Fig. 4: ROC with confidence intervals for the logistic regression. Left: for model # 4 in Table 4; AUC ≈ 0.80; 95 % confidence interval (0.62 0.91). Right: for model # 8 in Table 4; AUC ≈ 0.80; 95 % confidence interval (0.62 0.91).

Table 4: Examples of regression models: features used by the model, symbolic model description, optimal Operating Point (OP) on the ROC for a given model, and AUC estimates for the respective ROC.

Model index	Feature #	Regression model	Optimal OP		AUC		
			FPR	TPR	Min	Mean	Max
1	181, 136, 88	$1+x_1+x_2+x_3+x_1^2$	**0.3750**	**0.9167**	0.8229	0.6437	0.9286
2	181, 136, 88, 211	$1+x_1+x_2+x_3+x_4+x_1^2$	0.3750	0.8333	0.8125	0.6229	0.9149
3	181, 136	$1+x_1+x_2+x_1^2$	0.3125	0.8333	0.7969	0.6184	0.9063
4	181, 136, 211	$1+x_1+x_3+x_1^2$	0.1875	0.7083	0.7969	0.6241	0.9071
5	181, 20, 88	$1+x_1*x_2+x_2^2$	**0.2500**	**0.8333**	0.8307	0.6339	0.9396
6	181, 20, 88, 211	$1+x_1+x_4+x_1^2$	0.1875	0.7083	0.7969	0.6315	0.9093
7	181, 20	$1+x_1*x_2+x_2^2$	**0.2500**	**0.8333**	0.8307	0.6549	0.9336
8	181, 20, 211	$1+x_1+x_3+x_1^2$	0.1875	0.7083	0.7969	0.6196	0.9127
9	181, 20, 136	$1+x_1+x_3+x_1^2$	0.3125	0.8333	0.7969	0.6350	0.9141
10	181, 20, 136, 211	$1+x_1+x_4+x_1^2$	0.1875	0.7083	0.7969	0.6119	0.9010
11	181, 121, 136, 133	$1+x_2+x_4+x_4^2$	**0.2500**	**0.8333**	0.8151	0.6477	0.9211
12	181, 121, 136, 133, 211	$1+x_2+x_4+x_5+x_4^2$	0.1875	0.7500	0.8151	0.6309	0.9167

Table 5: Estimates of the regression coefficients, standard error SE, tStat, and pValue for these coefficients for regression models from Table 4.

Model index	Variable	Estimate of coefficient	SE	tStat	pValue
1	Intercept	58.615	18.365	3.1916	0.0014148
	x_1	-0.041377	0.015378	-2.6906	0.0071316
	x_2	-273.05	115.63	-2.3615	0.018202
	x_3	165.89	82.991	1.9989	0.045616
	x_1^2	8.1615e-06	3.4149e-06	2.39	0.01685
5	Intercept	-261.78	110.97	-2.3591	0.018321
	x_1x_2	-0.08107	0.028117	-2.8833	0.0039352
	x_2^2	-188.73	102.89	-1.8342	0.066625
7	Intercept	-261.78	110.97	-2.3591	0.018321
	x_1x_2	-0.08107	0.028117	-2.8833	0.0039352
	x_2^2	-188.73	102.89	-1.8342	0.066625
11	Intercept	61.581	22.969	2.681	0.0073398
	x_2	-0.0079267	0.0018617	-4.2577	2.0655e-05
	x_4	-0.067689	0.029635	-2.2841	0.022365
	x_4^2	2.4222e-05	9.7107e-06	2.4943	0.01262

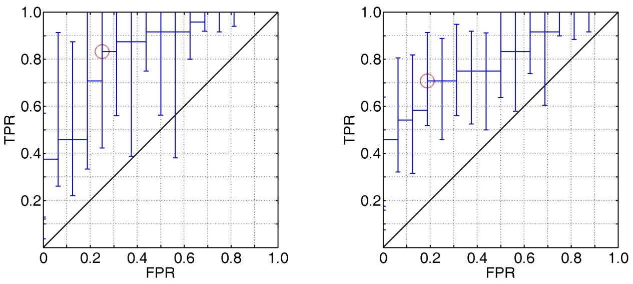

Fig. 5: ROC for the logistic regression. Left: for model # 5 in Table 4; AUC ≈ 0.83; 95 % confidence interval (0.63 0.94). Right: for model # 6 in Table 4; AUC ≈ 0.83; 95 % confidence interval (0.63 0.91).

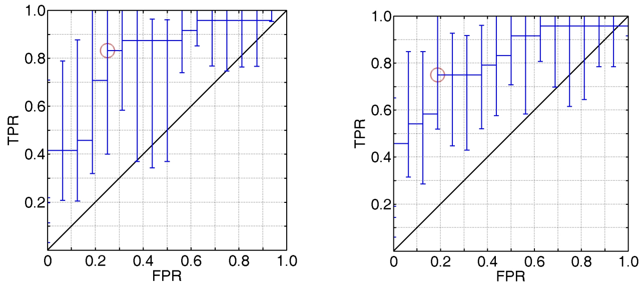

Fig. 6: ROC for the logistic regression. Left for model # 11 in Table 4; AUC ≈ 0.82; 95 % confidence interval (0.65 0.92). Right: for model # 12 in Table 4; AUC ≈ 0.82; 95 % confidence interval (0.63 0.92).

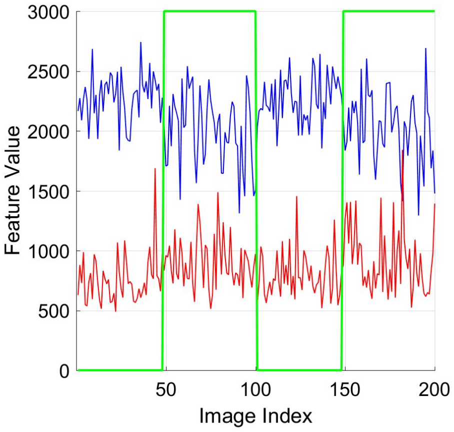

Fig. 7: Values of selected features in the database as a function of image index. Blue line represents the feature # 181; red line represents the feature # 211 (tumor circularity) multiplied by 200; green (rectangular) line shows the true label multiplied by 3000.

In order to get some feeling for the classification process look at Fig. 7 which shows the values of the contrast for vertical offset of 5 and horizontal offset of 4 (feature # 181) as well as tumor circularity (feature # 211) for all the images in the OASBUD database. For the purposes of the creation of this image, the hundred Rf1 scan plane images are sorted into 48 benign images followed by 52 malignant images. Subsequently, the hundred Rf2 scan plane images are sorted into 48 benign images followed by 52 malignant images. The green line in Fig. 7 represents the true labels of all the images in the database and forms a rectangular wave with upper and lower levels of **one** and **zero**. The selected contrast (feature # 181) and the tumor circularity (feature # 211) look like very noisy rescaled versions of the original rectangular wave.

As an example, Fig. 8 and Table 6 jointly illustrate the classification results for the regression model # 1 in Table 4. The first row of images in Fig. 8 shows two examples of benign lesion followed by three examples of malignant lesion taken from Rf1 scans of the database. The second row shows two examples of benign lesion followed by three examples of malignant lesion taken from Rf2 scans of the database. The images in Fig. 8 represent one quarter of the 40 images used in the testing set. All data needed for identification of images are given in Table 6. The windows containing the lesions have highly varying dimensions; however, in order to have a more regular figure we resized the images to the same height. This simplifies the figure, but may be misleading when one wants to compare malignant and benign cases. The reader might check the original sizes of images by looking into the database. Table 6 shows the results of classification of the images from Fig. 8 for several values of the threshold Th. Table 6 actually demonstrates how by decreasing the threshold Th introduced in Eqs. 2 and 3 we tend to increase the number of positive classifications and reduce the number of negative classifications, so in a sense it is a graphical illustration of the ROC.

Table 6: Lesion labels obtained from the model # 1 in Table 4 for the images in Fig 8. **One** denotes malignant, and **zero** denotes benign label.

Image index	Rf	True label	Threshold Th				
			0.1	0.3	0.5	0.7	0.9
53	Rf1	0	1	1	0	0	0
41	Rf1	0	1	1	1	1	0
96	Rf1	1	1	1	1	1	0
92	Rf1	1	1	1	1	1	1
65	Rf1	1	1	1	1	1	1
45	Rf2	0	1	1	1	1	0
43	Rf2	0	1	1	1	1	0
81	Rf2	1	1	1	1	0	0
60	Rf2	1	1	1	0	0	0
77	RF2	1	0	0	0	0	0

DISCUSSION

It can be seen from Fig. 2 that the decision trees do not give as good results as might be desired. No matter how we change the $MinLeaf$ parameter, the 0.632 bootstrap estimated ROC curve is distant from the ideal corner point of coordinates (0, 1). Furthermore, the inclusion of the most significant lesion shape parameter, that is tumor circularity, does not significantly improve the ROC curve. However, Zhou *et al.* (2015) indicated that this is the shape parameter that carries most of the information. The second best parameter is the standard deviation of degree. The definition of the standard deviation of degree given in (Zhou *et al.*, 2015) is based on the knowledge of the relationship between, say, the sth contour pixel and the $(s-k)$th and $(s+k)$th pixels. However, their paper does not specify the k, and our experiments did not give any convincing value, which would have a strong discriminative power. In any case, the corresponding wave for the deviation of degree similar to those in Fig. 7 seemed too noisy to be of much use.

In each pair of odd and even regression models in Table 4 both models have the same features except that the even model has an extra tumor circularity feature. Comparing their mean AUCs we observe that the addition of the tumor circularity either does not change the AUC or reduces it slightly. Similar general

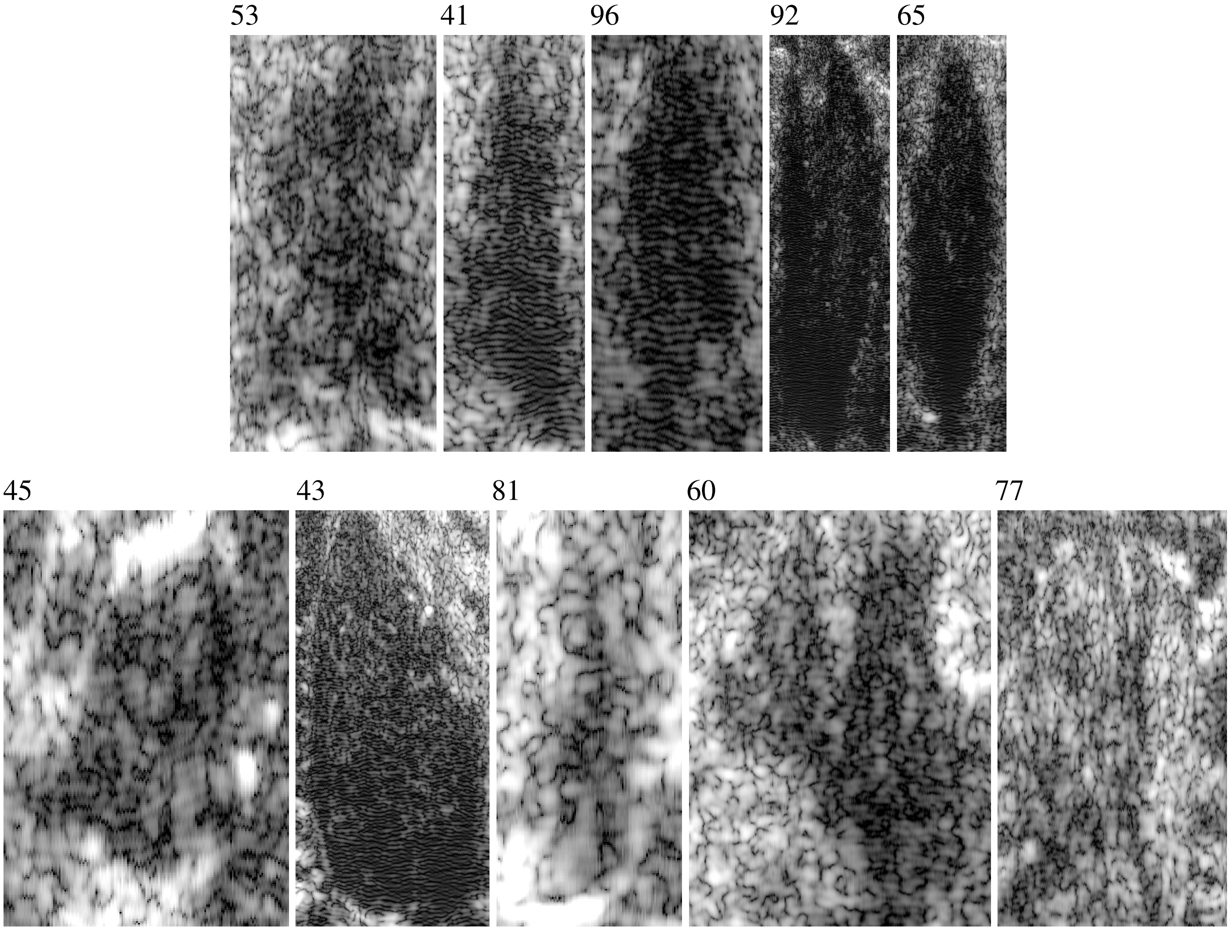

Fig. 8: Examples of windows used for classification to be considered jointly with their classification results in Table 6. The first row refers to images from the Rf1 scans in the database, and the second row refers to Rf2. All windows are resized to the same height exclusively for the purposes of creating this figure; in fact they are highly disparate. The numbers above the images are indices of images in the OASBUD database.

tendency occurs for the upper and lower bounds of the confidence intervals with few minor exceptions. The obvious conclusion is that the tumor circularity is in fact redundant.

Comparing regression models # 1 and 3 in Table 4 we see the reduction of the mean AUC from approx. 0.82 to 0.80. Similar reduction is observed when comparing models # 2 and 4. It means that in this case the use of the extra feature proved beneficial.

Comparing regression models # 5 and 7 we do not see any change in the mean AUC. Similar observation is valid for models # 6 and 8. However, comparing models # 5 and 6, or 7 and 8, we observe that the addition of the tumor circularity reduces the maximum pValue, which goes down from approx. 0.0666 to 0.0239. This means that by adding tumor circularity we get a more predictable model.

Comparing regression models # 7 and 9 we observe that the addition of the extra feature # 136 reduces the mean AUC. However, the regression model # 9 does not use the feature # 20 so it really is a 2-feature model. Comparing regression models # 8 and 10 we do not observe any change in the mean AUC. The regression model # 8 removes one feature, and model # 10 removes two features.

Comparing regression models # 3 and 11 we conclude that addition of extra feature # 121 and # 133 increases the mean AUC. However, model # 11 removes the features # 181 (contrast at offset (5, 4)) and # 136 (homogeneity at offset (4, 1)).

It worthwhile to observe that limiting the models to the first and second powers of the variables was established by the experimentation. Extending the models to the third power did not improve the AUCs.

The overall conclusion from reviewing Table 4 is that using stepwise generalized linear model we obtain the model which is best in some sense. The initial selection of the features for stepwise logistic regression was based on the inspection of the generated decision trees. But there are several decision trees obtained with various apparent ROCs. A good starting point is to select the features maintained by a single tree or several of them. In each case, the overwhelming number of features are rejected by the decision trees, and it does not make sense to consider them for the generation of regression model. It was observed that the selection of the features for a model is not critical and we get similar AUCs for several ROCs. When choosing the best model in Table 4 we have four best options which are bolded in this table and give AUC up to 0.83 with the 95 % confidence region (0.65 0.93).

The main steps in the evaluation of the classifier were model generation using the teaching set and subsequent testing the performance of the classifier by means of the testing set. The two sets were completely disjoint and selection of features could not bias the classification results. Trying to optimize both steps requires striking a balance between the usage of the feature vectors for building the model and for calculating the ROCs for the testing set. The assumed split into 160 vectors for teaching vs 40 for testing is an example of such a balance. As shown in Table 4 the obtained regression models are statistically reliable since the pValues in most cases do not exceed 0.05. At the same time lower boundary of 95 % confidence for most of the points on the ROC curve in Figs. 3 - 6 lies above the (0, 0) - (1, 1) diagonal in the presented diagrams. This means that we obtain results certainly above the random classification line represented by a (0, 0) - (1, 1) diagonal. Obviously, it would be desirable to have a bigger database so one could obtain smoother ROCs.

Further experiments with decision trees and stepwise logistic regression models showed that including all the six shape/contour features mentioned in the Shape Features section together with texture features resulted in a systematical rejection of most of the shape features leaving only texture features and tumor circularity in the model. In other words, the said shape/contour features turned out to be statistically insignificant when compared with the texture features and tumor circularity obtained from the raw RF signal.

An attempt was also made to use the Least Absolute Shrinkage and Selection Operator (LASSO), which returns fitted least-squares regression coefficients for linear models for feature vectors collected in a matrix X and desired labels collected in a vector Y. The features generated via GLCM are highly redundant, and one would hope that LASSO would identify and reject redundant features. The experiments conducted in Matlab confirmed this supposition. However, carrying the computation for features # 145,... , # 211, that is taking the last column and the top row in Table 1 we obtained LASSO results after 774 secs. And adding the fifth column and second row in Table 1 stretched this time to several hours. Taking into account all of the texture features seemed unrealistic. In this situation we undertook experimentation with generalized linear model and stepwise regression as described above.

The current paper obviously is somewhat parallel to papers using deep learning for classification of breast lesions (Byra *et al.*, 2018a). It is not clear which approach will turn out the best in the future. However, the method proposed in the current paper tries to explain the mechanism of classification in the possibly simplest way. The obtained results indicate that in fact two - three features can explain the classification. The deep learning with its multiplicity of features and thousands of teaching samples and thousands of weights which have to be adjusted seems to be excessively complicated. The ultimate comparison of various classifiers should be done by means of testing statistical hypotheses. In principle, this could be done but requires a larger number of samples (images). Currently available small medical databases allow one to build workable models but are insufficient for testing hypotheses that would compare performance of these models.

CONCLUSION

Comparison of the presented results of breast lesion classification with those obtained for raw US signals by means of maps of parameters of the homodyned K distribution (Byra *et al.*, 2016) indicates that both approaches give similar ROCs. Hence, the advantage of using GLCM-generated features is that this approach is significantly simpler, more homogeneous, and can be easily adjusted to a particular equipment or operating conditions. Inspection of Table 3 reveals that for various values of $MinLeaf$ and *acost*, a relatively small number of features are selected when the decision tree is generated for texture features obtained from the GLCM.

Using stepwise logistic regression models with features automatically pre-selected via the process of generating decision trees based on GLCM texture features and tumor circularity significantly improves the breast lesion classifier performance. For example, the three optimal operating points in Table 4 have coordinates (0.25, 0.83), and one point has coordinates (0.38, 0.92). These points are closer to the upper left corner (0, 1) of the square in right most diagram in Fig. 2, which shows the 0.632 bootstrap estimated pseudo-ROCs for the decision trees.

The other conclusion concerns the shape/contour features. The use of the shape/contour features of the lesion in addition to texture features proved to be of little effect when using stepwise logistic regression models. Experiments showed that similar observation is valid for the decision trees as well. This conclusion is conditioned on employing the upper half-mask of the lesion rather than full mask. The use of full mask involves indirect use of extra information related to the area of the acoustic shadow, which is difficult to obtain automatically and may require the help of a medical specialist; and this is not recommended since it would go against the purpose of our endeavor, which is to alleviate the work of the specialist and not create extra workload.

More future-oriented projects should aim at automatic detection and segmentation of the lesion more independent of the medical specialist.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Andrzej Nowicki from the Institute of Fundamental Technological Research in Warsaw, Poland for directing their attention to the OASBUD database and related publications.

REFERENCES

Adler, Lausen B (2009). Bootstrap estimated true and false positive rates and ROC curve. Computat Statistics Data Anal 53:718-29.

Alacam B, Yazici B, Bilgutay N (2003). Breast tissue characterization based on ultrasound RF echo modeling and tumor morphology. In: Proc 25th Annu Int Conf IEEE EMBS. pp. 1180-83.

Alacam B, Yazici B, Bilgutay N, Forsberg F, Piccoli C (2004). Breast tissue characterization using FARMA modeling of ultrasonic RF echo. Ultrasound Med Biol 10:1397-1407.

Alvarenga AV, Infantosi AFC, Pereira WCA, Azevedo CM (2012). Assessing the combined performance of texture and morphological parameters in distinguishing breast tumors in ultrasound images. Med Phys 39:7350-58.

Bingham NH, Fry JM (2010). Regression. Linear Models in Statistics. London: Springer.

Byra M, Nowicki A, Piotrzkowska-Wróblewska H, Dobruch-Sobczak K (2016). Classification of breast lesions using segmented quantitative ultrasound maps of homodyned K distribution parameters. Med Phys 43:5561-69.

Byra M, Dobruch-Sobczak K, Piotrzkowska-Wróblewska H, Nowicki A (2017). Added value of morphological features to breast lesion diagnosis in ultrasound, http://arxiv.org/abs/1706.01855

Byra M (2018). Discriminant analysis of neural style representations for breast lesion classification in ultrasound. Biocybernetics Biomed Eng 38:684-90.

Byra M, Sznajder T, Koržinek D, Piotrzkowska-Wróblewska H, Dobruch-Sobczak K, Nowicki A, Marasek K (2018a). Impact of ultrasound image reconstruction method on breast lesion classification with neural transfer learning, http://arxiv.org/abs/1804.02119v1

Chen D-R, Kuo W-J, Chang R-F, Moon WK, Lee CC (2002). Use of the bootstrap technique with small training sets for computer-aided diagnosis in breast ultrasound. Ultrasound Med Biol 28:897-902.

Cheng HD, Shan J, Ju W, Guo Y, Zhang L (2010). Automated breast cancer detection and classification using ultrasound images: a survey. Pattern Recogn 43:299-317.

Drukker K, Giger ML, Mendelson EB (2003). Computerized analysis of shadowing on breast ultrasound for improved lesion detection. Med Phys 30:1833-42.

Efron B, Tibshirani R (1998). An Introduction to the Bootstrap. Boca Raton: Chapman and Hall/CRC.

Gokhale S (2009). Ultrasound characterization of breast masses. Indian J Radiol Imaging. 19:242-47.

Gonzalez RC, Woods RE, Eddins SL (2009). Digital Image Processing Using MATLAB. Gatesmark Publishing.

Granchi S, Vannacci E, Biagi E, Masotti L (2015). Differentiation of breast lesions by use of hyperspace: hyper-spectral analysis for characterization in echography. Ultrasound Med Biol 41:1967-80.

Harvey P, Arger PH, Conant EF, Sehgal CM (2009). Differentiation of solid benign and malignant breast masses by quantitative analysis of ultrasound images. In: Proc IEEE Ultrasonics Symposium. Department of Computer Science, University of Copenhagen. pp. 530-33.

Hastie T, Tibshirani R, Friedman J (2009). The Elements of Statistical Learning. New York: Springer.

Kleinbaum DG, Klein M (2010). Logistic Regression. New York: Springer.

Kotsiantis SB (2013). Decision trees: a recent overview. Artif Intell Rev 39:261-83.

Lee H-W, Liu B-D, Hung KI-C, Lei S-F, Wang P-C, Yang T-L (2009). Breast tumor classification of ultrasound images using wavelet-based channel energy and imageJ. IEEE J Select Topics Image Proc 3:81-93.

Lizzi F, Astor M, Feleppa EJ, Shao M, Kalisz A (1997). Statistical framework for ultrasonic spectral parameter imaging. Ultrasound Med Biol 23:1371-82.

McCullagh P, Nelder JA (1989). Generalized Linear Models. London: Chapman & Hall/CRC.

Menon RV, Raha P, Chakrabarti I (2016). Classification of breast mass in ultrasound images using CAD: a survey. In: Proc Int Conf Systems in Medicine and Biology. IIT Kharagpur. EMBS, IEEE. pp. 31-35.

Minavathi M, Murali S, Dinesh MS (2012). Classification of mass in breast ultrasound images using image processing techniques. Int J Comput Applic 42:29-36.

Moon WK, Lo C-M, Cho N, Chang JM, Huang C-S, Chen J-H, Chang R-F (2013). Computer-aided diagnosis of breast masses using quantified BI-RADS findings. Computer Methods Programs Biomed 111:84-92.

Nemat H, Fehri H, Ahmadinejad N, Frangi AF, Gooya A (2018). Classification of breast lesions in ultrasonography using sparse logistic regression and morphology-based texture features. Med Phys 45:4113-24.

Nieniewski M, Zajączkowski P (2014). Real-time speckle reduction in ultrasound images by means of nonlinear coherent diffusion using GPU. In: Proc Int Conf Comput Vis Graphics. LNCS No. 8671, Springer. pp. 462-69.

Padilla PP, Bernardo DC, Encinas MAO, Marcos R, Castro AH, Palacios AL (2013). Ultrasound non-neoplastic breast lesions with posterior acoustic shadowing. http://dx.doi.org/10.1594/ecr2013/C-0058

Piotrzkowska-Wróblewska H, Nowicki A, Litniewski J, Gambin B, Dobruch-Sobczak K (2014). Breast carcinoma tissues characterization using statistics of ultrasonic backscatter. In: 7th Forum Acusticum 2014, 9 pages, http://www.ippt.pan.pl/en/staff/hpiotrzk

Piotrzkowska-Wróblewska H, Dobruch-Sobczak K, Byra M, Nowicki A (2017). Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. Med Phys 44:6105-09.

Pratiwi M, Alexander, Harefa J, Nanda S (2015). Mammograms classification using gray-level co-occurrence matrix and radial basis function neural network. Procedia Comput Science 59:83-91.

Sadeghi-Naini A, Suraweera H, Tran WT, Hadizad F, Bruni G, Rastegar RF, Curpen WT, Czarnota GJ (2017). Breast-lesion characterization using textural features of quantitative ultrasound parametric maps. Scientific Reports 7:1-10, http://www.nature.com/scientificreports

Uniyal N, Eskandri H, Abolmaesumi P, Sojoudi S, Gordon P, Warren L, Rohling RN, Salcudean SE, Moradi M (2015). Ultrasound RF time series for classification of breast lesions. IEEE Trans Med Imag 34:652-61.

Walach E, Kisilev P, Chevion D, Barkan E, Harary S, Hashaul S, Ben-Horesh A, Tzadok A (2013). A fully automatic lesion classification in breast ultrasound. In: Workshop Breast Image Analysis in conjunction with MICCAI 2013. Technical Report No 01/2013. Department of Computer Science, University of Copenhagen. pp. 98-105.

Zakeri FS, Behnam H, Ahmadinejad N (2012). Classification of benign and malignant breast masses based on shape and texture features in sonography images. J Med Syst 36:1621-27.

Zhou Z, Wu S, Chang K-J, Chen W-R, Chen Y-S, Kuo W-H, Lin C-C, Tsui P-H (2015). Classification of benign and malignant breast tumors in ultrasound images with posterior acoustic shadowing using half-contour features. J Med Biol Eng 35:178-87.

APPENDIX: OBTAINING THE UPPER HALF-MASK

The process of obtaining the upper half-mask and contour of a lesion includes the following steps:

1. Generation of a larger window containing the lesion and the corresponding contour as shown in Fig. 1(d). The motivation for the enlargement of the window is given in Database subsection.

2. Speckle removal in the larger window. There are many possibilities of speckle filtering. In the current paper one variant, that of nonlinear anisotropic diffusion was assumed following Nieniewski and Zajączkowski (2014) because of the availability of the software. All the parameters required were assumed as in their paper. The number of diffusion iterations was set to five. In fact, this number should be large enough to filter out the speckles and small enough not to filter out the lesion spicules (or minispicules). This is a tough requirement, but one can argue that any reasonable iteration number would do if used consistently.

3. Generation of a precise lesion contour. This step was executed by means of the Matlab command:

```
g = activecontour(f,mask,300,′Chan-Vese′,...
′Smoothingfactor′,3)
```

where `g` is the generated lesion contour; `f` is the image obtained from the anisotropic diffusion; and `mask` is the mask provided by the medical expert, 300 is the maximum number of iterations, `′Chan-Vese′` is the selected active contour method, and parameter 3 is the smoothing factor.

4. Automatic correction of the contour obtained in step 3. In almost all the cases the result from the step 3 is a single closed contour. It may happen, however, that the contour splits into a large contour and 1 - 2 small ones. The correction of the contour consists in removing those small parasitic contours. In fact, there were two such cases for the specified number of diffusion iterations and assumed active contour parameters.

5. Generation of the upper half-contour of the lesion. In principle, this step consisted in finding the leftmost and rightmost pixels on the contour obtained in step 4. In continuation we removed these extreme pixels and their closest neighbors in order to effectively break the contour into two parts. Subsequently we did morphological reconstruction of the upper-half in order to get rid of the lower part. Finally we completed the upper-half contour by means of the Bresenham algorithm generating a straight line between the end points of the upper half-contour.

6. Generating the mask of the upper-half of the lesion and cleaning the mask. The mask was generated by the morphological filling of the upper half-contour. An example of the mask obtained for images in Figs. 1(a)-(d) is shown on the left side of Fig. 9. It can be observed that this mask has a tiny "peninsula" at its low right corner. In some cases this "peninsula" can be significantly larger. Such a mask extension arises whenever the Bresenham line happens to be locally parallel to the mask boundary. In order to get rid of the "peninsula" we performed a sequence of morphological operations: `imerode` by 3×3 structuring element, `bwareaopen` with threshold 20 for removing all small spurious objects, and `imdilate` by 3×3 structuring element. This sequence of operations gave satisfactory masks for all the images in the OASBUD database. The result of cleaning the mask on the left side of Fig. 9 is shown on the right side of this figure.

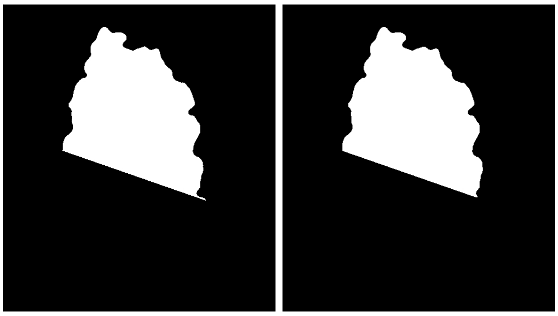

Fig. 9: Cleaning the lesion upper-half mask. Left: Original upper-half mask. Right: The same mask with removed "peninsula" in the lower right corner of the mask.

APPENDIX: ESSENTIAL MATLAB COMMANDS

Several Matlab commands employed by the authors are described below because their use is by no means obvious and involves specification of various options.

1. The command for obtaining B-mode images is:

```
envelope_image = 20 * log10(abs(hilbert(rf)))
```

It generates the `envelope_image` B-style image from the RF signal `rf` given in the database.

2. The command for calculating the GLCM is:

```
[GLCM] = graycomatrix(f,′NumLevels′,256,...
′offset′, [0 1],′Symmetric′,true)
```

where `f` is the input image (window), and `GLCM` is the Gray-Level Cooccurrence Matrix as defined in the book by Gonzalez *et al.* (2009). The set number of gray levels in the GLCM is 256. The exemplary `offset` is one pixel in one direction. The option `Symmetric,true` specifies that the function does not make difference depending on which pixel has which brightness. The first four features are calculated by means of the function `graycoprops`, as described in the book by Gonzalez *et al.* (2009). The last two features are calculated for the normalized GLCM, following the same book.

3. The command for tree generation is:

```
t1 = fitctree(d,g,′Prune′,′on′,...
′SplitCriterion′,′gdi′,′MinLeaf′,aleaf,...
′Cost′,[0 1;acost 0])
```

where `t1` is a data structure containing the tree, `d` contains feature vectors, `g` contains desired labels. The split criterion `gdi` is a Gini diversity index; `MinLeaf` denotes minimum number of observations admitted for a leaf (`aleaf`). The parameter `Cost` stands for a cost matrix, which is coded in accordance with Eq. 5.

4. The command for generating the regression model is:

```
mdl = stepwiseglm...
(TR_SET,Y,'linear','upper','poly222',...
'Distribution','binomial','link','logit',...
'Verbose',2),
```

where `mdl` is a data structure containing the generated linear model, `TR_SET` is the training set of 160 feature vectors (comp. Materials and Methods section), and `Y` is a vector of their desired labels. The parameter `linear` is a starting model for the stepwise regression; `poly222` is a specification of the upper model, in this case with three parameters, each of which can be of up to the second degree; `binomial` is a distribution of the classifier response variable (label); `logit` is a link function; and parameter `Verbose` of value 2 specifies what has to be printed out.

5. The command for predicting labels on the testing set is:

```
ypred = predict(mdl,TT_SET2)
```

in which `ypred` is a vector of the thresholds in the range (0, 1) indicating the probability of a success for a given test vector when sampling from a Bernoulli distribution. The `mdl` denotes the generalized linear model, and `TT_SET2` is the set of 40 test feature vectors (comp. Materials and Methods section).

6. The command for obtaining the ROC is:

```
[X,Y,T,AUC,OPTROCPT] = perfcurve(YTEST,...
ypred,'1','NBOOT',1000,'XVals','All')
```

In this command, `X,Y` denote coordinates of the points on the ROC curves; `T` is a vector of thresholds corresponding to the calculated `X,Y` points; `AUC` is the area under ROC curve; `OPTROCPT` are the coordinates of the optimal working point on the ROC curve. The `YTEST` denotes the vector of true class labels of the feature vectors in the testing set; and `ypred` denotes their predicted class labels. The 1 is a label for positive class. The 95 % confidence regions for points on the ROC curve are obtained by specifying the parameter `NBOOT` which means that the bootstrap with 1000 samples has to be used for vertical averaging. The parameter `Xvals` with value `All` indicates that all values of `ypred` have to be used.