Sonoelastomics for Breast Tumor Classification: A Radiomics Approach with Clustering-Based Feature Selection on Sonoelastography

Qi Zhang[1]*, Yang Xiao[2], Jingfeng Suo[1], Jun Shi[1], Jinhua Yu[3], Yi Guo[3], Yuanyuan Wang[3], Hairong Zheng[2]

1. Institute of Biomedical Engineering, Shanghai University, Shanghai, China.
2. Paul C. Lauterbur Research Center for Biomedical Imaging, Institute of Biomedical and Health Engineering, Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
3. Department of Electronic Engineering, Fudan University, Shanghai, China

***Author for Correspondence:**

Qi Zhang, Ph.D.
Associate Professor
Institute of Biomedical Engineering, Shanghai University, Shanghai, China.
E-mail: zhangq@shu.edu.cn; zhangq@t.shu.edu.cn
Tel: +86-21-66137256 Fax: +86-21-56338964
Address: Room 803, Xiangying Building, Shanghai University, No. 333, Nanchen Road, Shanghai, 200444, China

***Abstract*—A radiomics approach on sonoelastography, named "sonoelastomics," is proposed for classification of benign and malignant breast tumors. From sonoelastograms of breast tumors, a high throughput 364-dimensional feature set was calculated consisting of shape features, intensity statistics, gray level co-occurrence matrix texture features, and contourlet texture features, which quantified the shape, hardness, and hardness heterogeneity of a tumor. The high throughput features were then selected for feature reduction by using hierarchical clustering and three feature selection metrics. For a dataset containing 42 malignant and 75 benign tumors from 117 patients, seven selected sonoelastomic features achieved an area under a receiver operating characteristic curve of 0.917, an accuracy of 88.0%, a sensitivity of 85.7%, and a specificity of 89.3% in a validation set via the leave-one-out cross validation, demonstrating superiority over the principal component analysis, deep polynomial networks, and manually selected features. The sonoelastomic features are valuable for breast tumor differentiation.**

Keywords-Radiomics; sonoelastography; breast tumor; classification; feature selection; hierarchical clustering

INTRODUCTION

Ultrasound elastography or sonoelastography has emerged as a valuable tool for breast tumor characterization by depicting tissue hardness on color images (Barr et al. 2015, Zhang et al. 2014). Malignant and benign tumors have different color patterns on sonoelastography due to their different hardness distributions. There are mainly two categories of sonoelastography, strain elastography (Kadour and Noble 2009, Ophir et al. 1991) and shear wave elastography (Bercoff et al. 2004, Nightingale et al. 2003). Strain elastography is easy to use and provides elasticity images in a manner similar to palpation (Shiina et al. 2015). Many manufacturers produce medical ultrasound devices with a strain elastography function (Shiina et al. 2015). Considering its wider and wider availability, the present study is focused on strain elastography.

In clinical practice of strain elastography, the Tsukuba score is usually used for qualitative assessment of breast tumors, which is a five-point scale that visually grades the hardness of a mass (Itoh et al. 2006). Ten-point grading (Zhi et al. 2013), three-point grading (Kim et al. 2015) and another five-point grading (Alhabshi et al. 2013) are also employed. However, these grading methods suffer from considerable inter-observer variability because of its subjective and qualitative description of lesion hardness (Yoon et al. 2011).

Quantitative assessment has been proposed to provide less subjective and less operator dependent descriptions. It usually measures the ratio of the strain in fat or gland to the strain in a tumor, i.e., fat to lesion strain ratio or gland to lesion strain ratio (Cho et al. 2010, Fausto et al. 2015, Zhao et al. 2012, Zhou et al. 2014), or the ratio of the hard area within a tumor to the area of the entire tumor (i.e., area ratio) (Zhang et al. 2014). These ratios were proposed based on the fact that malignant breast tumors are usually harder than benign tumors. A feature related to tumor shape was also derived as the ratio of the lesion size on elastography to the

B-mode size (i.e., size ratio) (Alhabshi et al. 2013, Barr et al. 2015). However, these few descriptors have attained limited diagnostic performance, probably because they only focus on a certain aspect of the tumor hardness or shape while neglecting other useful information such as the tumor heterogeneity. Breast tumor is a heterogeneous tissue with intratumoral regional variations in proliferation, cell death, metabolic activity, vascular structure and other factors (Asselin et al. 2012, Zhang et al. 2015). The heterogeneity is also a pattern trait of malignancy (Chaddad et al. 2015, Zhang et al. 2015). Thus, the tumor shape, hardness, and heterogeneity should all be taken into consideration in breast tumor classification.

Recent advances in machine learning algorithms allow for more objective and precise quantitative imaging descriptors, which could comprehensively evaluate breast tumor intensity, shape and texture and could potentially be used as noninvasive biomarkers for discrimination between malignant and benign tumors (Venkatesh et al. 2015). Radiomics refers to the extraction and analysis of a large number of quantitative features with high throughput from medical images (Aerts et al. 2014, Kumar et al. 2012, Lambin et al. 2012). Radiomics have been increasingly used in computer tomography, magnetic resonance imaging, and positron emission tomography (Gillies et al. 2015, Huang et al. 2016, Vallières et al. 2015), but seldom employed in ultrasonography. In this paper, we propose using a radiomics approach on sonoelastography for breast tumor classification, and thus we name the approach as "sonoelastomics." The high throughput features are then selected for feature reduction by using hierarchical clustering (HC). We hypothesize that the sonoelastomic features capture distinct differences of breast tumors and may have discriminative ability for tumor classification.

64 # MATERIALS AND METHODS

65 ***Image Acquisition, Hardness Retrieval and Image Segmentation***

66 Ethical approval was obtained and the informed consent requirement was waved for this
67 retrospective study. A sonoelastography dataset containing 117 patients with 117 breast tumors
68 (42 malignant and 75 benign) was used in the study. The elastograms were acquired before
69 tumor biopsy using the HI VISION Preirus system (Hitachi Medical System, Tokyo, Japan)
70 equipped with a 5-13 MHz linear array probe. All tumors were subjected to core biopsy or fine
71 needle aspiration cytology for histopathologic diagnosis as the gold standard. For examining
72 the repeatability of elastography, we acquired two images from each of 110 tumors at two
73 scanning planes or in a time interval of around 10 s, while there was only one image acquired
74 for each of the remaining 7 tumors.

75 The Hitachi Preirus elastography system provides dual-modality visualization in a full
76 screen (Fig. 1a), where the right part is a grayscale B-mode image, and the left part is a
77 composite color RGB image displayed as a translucent color elastographic image
78 superimposed on the grayscale B-mode image. Therefore, a pure color elastogram was
79 obtained by subtracting the B-mode image from the composite image, but still in RGB format
80 (Fig. 1b) (Zhang et al. 2016, Zhang et al. 2015, Zhang et al. 2014). The hardness distribution
81 was then retrieved by computing the hue (H) values from the pure elastogram (Zhang et al.
82 2014):

83 $$H = \begin{cases} H0, \text{if } B \le G \\ 1 - H0, \text{if } B > G \end{cases} \tag{1}$$

84 $$H0 = \frac{1}{2\pi} \cos^{-1} \left\{ \frac{2R - G - B}{2\sqrt{(R-G)^2 + (R-B)(G-B)}} \right\} \tag{2}$$

where *R*, *G* and *B* were three color values of a pixel in the pure elastogram. The Hitachi elastography system only uses 5/6 part of the full hue scale, namely from red to blue (color bar on Fig. 1a), but without colors such as purple and purplish-red that are covered in the remaining 1/6 part. Therefore, the *H*-value calculated from (1) quantifies tissue hardness and ranges from 0 (red, softest) to 5/6 (blue, hardest), depicted as the grayscale image in Fig. 1e. There are missing areas without hardness information on elastograms, which appear as black holes or shades (Fig. 1a and Fig. 1b). The pixels in these areas have invalid hue values and were automatically detected and excluded from further analysis (Fig. 1e).

An automated image segmentation method using the Chan-Vese level sets was applied to B-mode images to detect tumor boundaries, followed by a morphologic closing operation (Zhang et al. 2015, Zhang et al. 2014). The tumor boundaries detected on B-mode images (Fig. 1c) were then mapped to the retrieved elastograms (Fig. 1e) to specify the regions of interest.

Feature Generation

Four categories of features were calculated, namely the shape features, intensity statistics, gray level co-occurrence matrix (GLCM) texture features, and contourlet texture features.

The shape features quantified the morphology of tumors. They included the area, convex area, perimeter, equivalent diameter, long-and short-axis lengths, orientation, solidity, eccentricity, as well as the mean, median and maximal thicknesses, and the mean, median and maximal widths.

The intensity statistics quantified the intensity distributions on the elastograms and were calculated from the hue values (i.e., hardness) within a tumor, including a variety of first order statistics such as the mean, standard deviation, coefficient of variance (Cov), skewness, kurtosis, entropy of histogram (EtH), area ratio, and combined area ratio (CAR), and several

percentiles (Zhang et al. 2014, Zhang et al. 2015). Other features included the statistics outside a tumor, and ratios of statistics within a tumor to those outside a tumor.

Texture features were then calculated from the GLCM (Haralick and Shanmugam 1973). The GLCM was normalized to get the joint conditional probability density function, from which the texture features based on GLCM were derived, including the energy, contrast, homogeneity and entropy of GLCM (Zhang et al. 2014). To achieve a more efficient representation of the texture, the hue image was requantized to 8 intensities and hence, the size of GLCM was 8 × 8. In our practice, the GLCM was calculated at a distance of 1, 2, 3, 4 and 8 pixels and a direction of 0º, 45º, 90º and 135º. The GLCM-based texture features were averaged over the four directions (Zhang et al. 2014).

Texture features were also extracted based on the contourlet transform, which was conducted to decompose an elastogram into multiscale bandpass (BP) bands and lowpass (LP) bands (Do and Vetterli 2005, Zhang et al. 2015). Each BP band was further decomposed into multi-directional subbands (Do and Vetterli 2005, Zhang et al. 2015). We calculated texture features from the LP and BP contourlet bands, respectively. The LP band is equivalent to a blurred image after downsampling the original image, and thus the aforementioned intensity statistics and GLCM features were naturally derived from this band as its corresponding texture features. The BP band involves the edge information in the original image, and there are three methods for calculating its texture features: (a) The intensity statistics and GLCM features were directly computed from the BP band rather than the directional subbands. We named this method as the direct (DIR) method. (b) The intensity statistics and GLCM features were first computed from each directional subbands and then averaged across all directions, hereafter referred to as the subband averaging method. (c) A new series of subband signals were reconstructed by using directional filter banks, and the intensity statistics and GLCM

features were derived from these reconstructed signals as described in (Zhang et al. 2015). We named this method as the subband reconstruction averaging (SRA) method.

In total, there were 364 features, consisting of 15 shape features, 51 intensity statistics, 25 GLCM texture features, and 273 contourlet texture features.

Hierarchical Clustering and Heat Map Rendering

Hierarchical clustering (HC) has been widely used in gene expression data, specifically for genomics (Bar-Joseph et al. 2001, Golub et al. 1999). HC groups data over a variety of scales by creating an agglomerative cluster tree, namely a multilevel hierarchy where clusters at one level are joined as clusters at the next level. Here, we applied HC to sonoelastomics, rather than genomics, for exploring intrinsic patterns in sonoelastograms.

Let $X \in R^{m \times n}$ be a data matrix with m features and n samples, we performed HC along both rows and columns of the matrix. Specifically as shown in Fig. 2, the HC linked pairs of objects (rows or columns of X) that were close together into binary clusters, i.e., clusters made up of two objects. Here the distance between two objects was quantified by the Pearson correlation distance measure (Bar-Joseph et al. 2001, Golub et al. 1999). Subsequently, the HC linked these newly formed clusters to each other and to other objects so as to create larger clusters until all the objects in X were joined in a hierarchical tree. The HC first linked pairs of rows as the objects to form a hierarchical tree of m features, and then linked pairs of columns as the objects to form a hierarchical tree of n samples.

The distribution of each feature $x \in R^{1 \times n}$ was quantified by using the Z-score:

$$Z\text{-score} = \frac{x - \bar{x}}{\sigma} \tag{3}$$

where $\bar{x}$ and σ denoted the mean and standard deviation of a feature on all n samples. Z-score was then rendered as a heat map using a pseudo color map, together with the cluster trees generated from HC along both row and column directions (Fig. 2).

Feature Selection from Clusters and Classification

Features were selected from the high-dimensional feature set for feature reduction by using the clusters derived from HC along rows. Suppose we had obtained C clusters by performing HC along the rows of X. We then selected one typical feature from each cluster according to one of the following three metrics.

(a) We randomly distributed two images acquired from a same tumor (110/117) to two groups, and then computed the correlation coefficient (R) of each feature to measure its repeatability between two groups.

(b) The P-value of the independent two-sample t-test was yielded to examine the difference of each feature between benign and malignant tumors.

(c) The square root of Fisher inter-intra class variance ratio (F_v) was also adopted to further quantify the difference (Zhang et al. 2014):

$$F_v = (\bar{x}_0 - \bar{x}_1) / \sqrt{(\sigma_0^2 + \sigma_1^2)} \quad (4)$$

where the subscripts 0 and 1 represented benign and malignant classes, respectively.

Based on the three metrics, a typical feature with the largest R-value, largest absolute F_v-value or smallest P-value should be selected from a cluster. It is worth noting that the R-value is an unsupervised metric without use of class labels and the F_v-value and the P-value are supervised metrics.

The leave-one-out cross validation using the proposed feature selection method and the support vector machine (SVM) classifier was performed on 117 images, one image for one tumor, to assess the sensitivity (SEN), specificity (SPC), accuracy (ACC) and Youden's index (YI = SEN+SPC-1) of the classification. The leave-one-out cross-validation involved using a single tumor as the validation (test) set of the feature selection and classification and the remaining tumors as the training set, and this was repeated such that each tumor was used once as the validation set. Furthermore, on both the training and validation sets, a receiver operating characteristic (ROC) curve was derived by tuning the thresholds of cancer likelihood. Cancer likelihood was a posterior probability between 0 and 1, and it was calculated with Platt's algorithm by mapping the distance of each sample to the decision boundary of the classifier using a sigmoid function (Platt 1999, Uniyal et al. 2015, Zhang et al. 2016). For each training set containing 116 samples, the threshold of cancer likelihood was tuned from 0 to 1 to get various classification results (i.e., SEN and SPC), yield an ROC curve, and calculate an area under the ROC curve (AUC). Each validation set only contained one sample, and 117 validation sets were combined to include all 117 samples so that the threshold of cancer likelihood was tuned to derive one ROC curve for the validation sets and get the AUC value.

EXPERIMENTS AND RESULTS

We first clustered the 117 images into two groups along the columns (i.e., samples) of the data matrix X to evaluate the classification performance of the purely unsupervised learning. We adjusted the cluster number C along the rows (i.e., features) from 2 to 15 to search for the optimal parameter for our radiomics classification scheme. Three metrics used in feature selection were evaluated in terms of classification performance.

Our scheme was compared with several methods: a) a method using all features as the input of an SVM classifier; b) a classic method using the principal component analysis (PCA) for

feature reduction and SVM for classification (named PCA-SVM); c) a method using eight recently reported, manually selected features (Zhang et al. 2015) as the input of SVM (named ManualSel); and d) a recently proposed deep learning algorithm, namely the deep polynomial network (DPN) (Livni et al. 2013). The DPN is a new type of multi-layer neural networks, in which the output of each node at the first and last layers is linear weighted sum of its input variables, and the output of each node at the intermediate layers is a quadratic function of its inputs. For the last layer, Livni et al. (Livni et al. 2013) utilizes stochastic gradient descent to train a linear classifier, using an L_2-regularized hinge loss (denoted as DPN-Hinge). For more comprehensive comparisons, we also modified the classifier in DPN to linear SVM (DPN-SVM) and fisher classifier (DPN-Fisher). All the parameters in the compared methods were set empirically to achieve best performance.

Heat Map and Cluster Trees

The *Z*-score is illustrated as a heat map in Fig. 2, and the cluster trees obtained from HC are depicted on the left and top of the heat map. The 364 rows (i.e., features) were agglomerated into 7 clusters, and 117 columns (i.e., samples) into 2 groups. The difference between Group I (79 samples) and Group II (38 samples) is visually distinct on abundant radiomics features. There was a significant difference of benign and malignant tumor proportions between the two groups obtained from HC ($P < 0.001$; χ^2 test), implying the two groups might well represent benign and malignant tumors. The clustering yielded an SEN of 52.4% (22/42), an SPC of 78.7% (59/75), an ACC of 69.3% (81/117), and a YI of 31.0%.

Among 364 features, 287 features exhibited significant differences between benign and malignant tumors ($P < 0.05$; t-test), and 174 features exhibited extremely significant differences ($P < 0.0001$).

221 ***Typical Features Selected from Clusters***

222 Table 1 shows typical features selected from 7 clusters when using the F_v-metric. One
223 typical feature was automatically selected from each of the 7 clusters with the largest absolute
224 F_v-value. Among the 7 selected features, there were two shape features (Eccentricity and
225 Solidity), one intensity feature (EtH), and four contourlet texture features.

226 Among all 364 features, the contourlet feature Median-SRA1 i.e., the median at the first
227 contourlet level using the subband reconstruction averaging method, had the largest absolute
228 F_v-value (-1.362) and also had a very large R-value (0.821) and an extremely small P-value
229 (6.31×10^{-17}) (Table 1). The average Median-SRA1 value was 0.0020±0.0003 in benign tumors
230 and 0.0025±0.0002 in malignant tumors, suggesting that the malignant tumors were more
231 heterogeneous on hardness and the heterogeneity could help for classification.

232 The intensity feature Mean, i.e., the mean hardness within a tumor, also had a high
233 discriminative ability (0.470±0.079 in benignancy and 0.594±0.061 in malignancy; P =
234 1.84×10^{-14}; F_v = -1.238; R = 0.810), suggesting that benign tumors were softer than malignant
235 tumors, which was in agreement with clinical findings. However, Mean was not a typical
236 feature selected by HC. This was because Mean was grouped into the same cluster where
237 Median-SRA1 joined, and its P-value, F_v-value and R-value were all weaker than those of
238 Median-SRA1. These results indicate that the hardness within a tumor and its heterogeneity
239 may be both valuable for tumor discrimination, and the heterogeneity may have a stronger
240 discriminative power.

241 The quantitative results are in accordance with the visual observation in Fig. 3 and Fig. 1,
242 where the malignant tumor is predominantly blue and heterogeneously mixed with green,

yellow and red (Fig. 3a), and the benign tumors are homogeneously shaded in green (Fig. 3c and Fig. 1d).

Classification Performance with Various Clusters and Three Feature Selection Metrics

Fig. 4 shows ACC and YI of our classification scheme in the validation set when tuning clusters numbers from 2 to 15. The F_v-metric achieved the best ACC (88.0%) and YI (75.0%) when C = 7. The P-metric yielded a high ACC (87.2%) and a high YI (72.7%) when C = 7. The unsupervised R-metric also obtained satisfactory results when C = 10, with an ACC of 87.2% and a YI of 72.7%. It should be noted that when using the R-metric, there was no need to know the class labels, and the features were selected in an unsupervised way by combining HC and repeatability examination.

Fig. 5 shows typical samples of breast tumors that were correctly classified with our method using F_v-metric, called Cluster-Fv, when C = 7. The malignant tumors shown in Fig. 5a-5d and benign tumors shown in Fig. 5k-5m can be easily interpreted and correctly classified by human observer or computer, because these malignant tumors appear predominantly blue indicating very stiff tissues and these benign ones are covered by green representing softer tissues. However, the malignant tumors shown in Fig. 5e-5j and benign tumors shown in Fig. 5n-5t are much more difficult to be interpreted and classified, because they both present blue and green staggered colors and thus are borderline cases. Especially in Fig. 5i-5j, there is a large portion of green inside a tumor, and in Fig. 5r and 5t, there is a large portion of blue inside, which can easily lead to misclassification when only considering the general hardness of tumors. The malignant borderline cases (Fig. 5e-5j) appear a color pattern more heterogeneous than the benign ones (Fig. 5n-5t). This detailed information is successfully captured by the texture features selected in the Cluster-Fv method, contributing to correct classification.

Comparisons with Classic Computerized Methods and Deep Learning Algorithms

As enumerated in Table 2, when using all features as the input of SVM, the AUC, ACC and YI were 0.811, 76.1% and 47.0% in the validation set, respectively. When using PCA for feature reduction, the AUC increased to 0.887, ACC to 85.5% and YI to 64.8%; when using manually selected features, the AUC increased to 0.890, ACC to 84.6% and YI to 64.5%. DPN-Hinge adopted a stochastic algorithm, and thus its results were averaged across 50 times of experiments. DPN-Hinge achieved the best SPC (94.2%±2.3%) among all methods, as well as a high ACC (87.2%±1.0%); however, its SEN was only 74.7%±2.2% and AUC was 0.870±0.005. DPN-Fisher yielded a good AUC (0.889) and SPC (89.3%), but its SEN (78.6%) and YI (67.9%) were not very high. DPN-SVM was worse than DPN-Hinge and DPN-Fisher in terms of most indices.

Our methods using three metrics are denoted as Cluster-R, Cluster-Fv, and Cluster-P. Cluster-Fv attained the highest values of AUC (0.917), SEN (85.7%), ACC (88.0%) and YI (75%) among all methods (Table 2), as well as a high SPC (89.3%). The ROC curves depicted in Fig. 6 further demonstrates the superiority of Cluster-Fv over both the classic and deep learning methods. Cluster-P attained a second large AUC-value (0.897), and fairly high ACC- and YI-values. Cluster-R achieved a reasonably high AUC of 0.885, indicating the unsupervised learning may also capture the intrinsic patterns on sonoelastograms.

Comparisons with Clinical Methods

Table 3 lists the classification performance in representative clinical publications. Because all 12 previous studies were conducted without cross-validation, we also list the results of Cluster-Fv without cross-validation for fair comparison. The AUCs across these studies ranged from 0.669 to 0.960, and the accuracies were between 69.3% and 95.4%. Our method yielded the second largest AUC (0.937) and the third largest accuracy (91.5%). It was superior to all

qualitative grading methods and all but one quantitative method (Zhang et al. 2014) in terms of AUC. The proportion of malignant tumors in our study is 35.9%, which was more balanced than the proportion in Kim et al. 2015 (10.1%) probably resulting in their over-estimation of the classification accuracy. Our dataset containing one tumor for each patient was also more appropriate for yielding reliable results, while in Zhang et al. 2014, the tumor number (145) was much more than the patient number (104), indicating the tumors were not independent and it might lead to biased results.

DISCUSSION AND CONCLUSIONS

The most important contribution of this work is to propose a quantitative radiomics approach on sonoelastography to breast tumor feature selection and classification. Specifically, high throughput features are generated from sonoelastography, and a feature subset is selected from the high-dimensional feature pool with hierarchical clustering and one of three metrics.

The proposed radiomics method for feature selection and tumor classification needs to be evaluated on an independent validation cohort. Furthermore, it should be elucidated whether the radiomics features (the high-dimensional or the selected) have prognostic power and could potentially be used as prognostic biomarkers for monitoring the development and progression of breast cancer or its response to therapy. These features are also expected to have diagnostic and prognostic power for other tumors or diseases. Moreover, the relationship between the sonoelastomic features of tumors and their underlying gene-expression patterns needs to be discerned by combining radiomics and genomics (Jamshidi et al. 2015). The radiomics features are also expected to be incorporated with laboratory parameters from blood tests for enhanced diagnosis and prognosis (Huang et al. 2016).

Along the columns (samples) of the data matrix, HC agglomerated the samples into various numbers of groups at various levels, not only two groups for possibly representing benignancy and malignancy. For example, both Groups I and II in Fig. 2 are composed of two sub-groups, which also consist of smaller sub-groups. These sub-groups may represent tumor sub-types, such as invasive ductal carcinoma and ductal carcinoma in situ in malignant tumors, and fibroadenoma and fibrocystic change in benign tumors. Some benign sub-types may be very close to malignant ones, making it difficult to discriminate between them. The radiomics approach using HC at various levels may be possible for differentiating tumor sub-types, which may potentially contribute to more precise diagnosis for personalized medicine. However, this hypothesis needs to be validated with a larger cohort.

When using all features or DPN algorithms, the classification results on the training set are much better than those on the validation set (Table 2), indicating that using the raw features without feature reduction or the deep learning methods may lead to over-fitting of the classification models. This fact is due to the small sample size (117) compared with the large feature dimensionality (364). Here we propose a radiomics approach using hierarchical clustering for feature selection, which can effectively suppress over-fitting and result in classification indices in the validation set as high as in the training set.

Although the *R*-metric did not use any information about class labels, in general, features with higher repeatability (*R*-values) showed higher absolute F_v-values and lower *P*-values, as well as larger classification indices. This is possibly due to reduced amount of noise in these more repeatable and stable features (Aerts et al. 2014). The *R*-metric may be also helpful for future studies in the clinical setting, where a large amount of images are available but only a few are labeled.

This study was focused on sonoelastomics, and the features on other modalities were not included except the shape features derived from B-mode ultrasound. We will combine features from B-mode, Doppler, and elastograms for multiple ultrasonic modality analysis using radiomics, and hence this extended analysis will be termed “ultrasonomics” or “sonomics.” In addition, our study was performed on one type of sonoelastography, and other types such as shear wave elastography should be compared in the future.

In this work, we did not use other categories of texture features such as gray-level run-length matrix, gray-level size zone matrix, and neighborhood gray-tone difference matrix (Vallières et al. 2015). It is due to the reason that among these matrix-based texture features, the GLCM is probably the most famous and prevailing method. It has been more familiar to medical community than other methods, and thus it may be more easily accepted by medical community. There are two methods for calculating GLCM features: one is averaging the textures over four directions, and the other is using a single matrix accumulating all co-occurrence measurements from all directions (Hatt et al. 2015, Vallières et al. 2015). We only employed the first method since it was more widely used. Nevertheless in future studies, we will include more categories of texture features and different methods of GLCM generation for increasing the throughput of features and further ameliorating the classification.

In conclusion, we propose using a radiomics approach on sonoelastograms, termed sonoelastomics, for generating high throughput quantitative features, from which a few typical features can be selected with the hierarchical clustering. The selected features can capture distinct differences between benign and malignant breast tumors and are valuable for breast tumor discrimination.

Acknowledgments

The work was supported by the National Science Foundation of China (No. 61671281, 61401267, 61302039, 81371560, 81627804 and 61471231) and the Chenguang Project from Shanghai Education Committee (No. 11CG45). We thank anonymous reviewers for their insightful and useful comments.

References

Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Cavalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5:4006.

Alhabshi SMI, Rahmat K, Halim NA, Aziz S, Radhika S, Gan GC, Vijayananthan A, Westerhout CJ, Mohd-Shah MN, Jaszle S, Semi-quantitative and qualitative assessment of breast ultrasound elastography in differentiating between malignant and benign lesions. Ultrasound Med Biol 2013;39:568-78.

Asselin MC, O'Connor JPB, Boellaard R, Thacker NA, Jackson A, Quantifying heterogeneity in human tumours using MRI and PET. Eur J Cancer 2012;48:447-55.

Bar-Joseph Z, Gifford DK, Jaakkola TS, Fast optimal leaf ordering for hierarchical clustering. Bioinformatics 2001;17 Suppl 1:S22-9.

Barr RG, Nakashima K, Amy D, Cosgrove D, Farrokh A, Schafer F, Bamber JC, Castera L, Choi BI, Chou YH, Dietrich CF, Ding H, Ferraioli G, Filice C, Friedrich-Rust M, Hall TJ, Nightingale KR, Palmeri ML, Shiina T, Suzuki S, Sporea I, Wilson S, Kudo M, Wfumb Guidelines and Recommendations for Clinical Use of Ultrasound Elastography: Part 2: Breast. Ultrasound Med Biol 2015;41:1148-60.

Bercoff J, Tanter M, Fink M, Supersonic shear imaging: a new technique for soft tissue elasticity mapping. IEEE Trans Ultrason Ferroelectr Freq Control 2004;51:396-409.

Chaddad A, Zinn PO, Colen RR. Radiomics texture feature extraction for characterizing GBM phenotypes using GLCM. Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on: IEEE, 2015. pp. 84-87.

Cho N, Moon WK, Kim HY, Chang JM, Park SH, Lyou CY, Sonoelastographic strain index for differentiation of benign and malignant nonpalpable breast masses. J Ultras Med 2010;29:1-7.

Do MN, Vetterli M, The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans Image Process 2005;14:2091-106.

Fausto A, Rubello D, Carboni A, Mastellari P, Chondrogiannis S, Volterrani L, Clinical value of relative quantification ultrasound elastography in characterizing breast tumors. Biomed Pharmacother 2015;75:88-92.

Gillies RJ, Kinahan PE, Hricak H, Radiomics: images are more than pictures, they are data. Radiology 2015;278:563-77.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.

Haralick RM, Shanmugam K, Textural features for image classification. IEEE Transactions on systems, man, and cybernetics 1973:610-21.

Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest CC, Groheux D, Hindié E, Martineau A, Pradier O, Hustinx R, 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort. J Nucl Med 2015;56:38-44.

Huang YQ, Liang CH, He L, Tian J, Liang CS, Chen X, Ma ZL, Liu ZY, Development and Validation of a Radiomics Nomogram for Preoperative Prediction of Lymph Node Metastasis in Colorectal Cancer. J Clin Oncol 2016;34:2157-64.

Itoh A, Ueno E, Tohno E, Kamma H, Takahashi H, Shiina T, Yamakawa M, Matsumura T, Breast disease: Clinical application of US elastography for diagnosis. Radiology 2006;239:341-50.

Jamshidi N, Jonasch E, Zapala M, Korn RL, Brooks JD, Ljungberg B, Kuo MD, The radiogenomic risk score stratifies outcomes in a renal cell cancer phase 2 clinical trial. Eur Radiol 2015:1-10.

Kadour M, Noble JA, Assisted-freehand ultrasound elasticity imaging. IEEE Trans Ultrason Ferroelectr Freq Control 2009;56:36-43.

Kim S-Y, Park JS, Koo HR, Combined use of ultrasound elastography and B-mode sonography for differentiation of benign and malignant circumscribed breast masses. J Ultras Med 2015;34:1951-59.

Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ, Radiomics: the process and the challenges. Magn Reson Imaging 2012;30:1234-48.

Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ, Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48:441-6.

Livni R, Shalev-Shwartz S, Shamir O, An algorithm for training polynomial networks. arXiv preprint arXiv:1304.7045 2013.

Nightingale K, McAleavey S, Trahey G, Shear-wave generation using acoustic radiation force: in vivo and ex vivo results. Ultrasound Med Biol 2003;29:1715-23.

Ophir J, Céspedes I, Ponnekanti H, Yazdi Y, Li X, Elastography: A Quantitative Method for Imaging the Elasticity of Biological Tissues. Ultrasonic Imaging 1991;13:111-34.

Platt J, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 1999;10:61-74.

Shiina T, Nightingale KR, Palmeri ML, Hall TJ, Bamber JC, Barr RG, Castera L, Choi BI, Chou YH, Cosgrove D, Dietrich CF, Ding H, Amy D, Farrokh A, Ferraioli G, Filice C, Friedrich-Rust M, Nakashima K, Schafer F, Sporea I, Suzuki S, Wilson S, Kudo M, Wfumb Guidelines and Recommendations for Clinical Use of Ultrasound Elastography: Part 1: Basic Principles and Terminology. Ultrasound Med Biol 2015;41:1126-47.

Uniyal N, Eskandari H, Abolmaesumi P, Sojoudi S, Gordon P, Warren L, Rohling RN, Salcudean SE, Moradi M, Ultrasound RF time series for classification of breast lesions. IEEE T Med Imaging 2015;34:652-61.

Vallières M, Freeman C, Skamene S, El Naqa I, A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol 2015;60:5471.

Venkatesh SS, Levenback BJ, Sultan LR, Bouzghar G, Sehgal CM, Going beyond a First Reader: A Machine Learning Methodology for Optimizing Cost and Performance in Breast Ultrasound Diagnosis. Ultrasound Med Biol 2015;41:3148-62.

Yoon JH, Kim MH, Kim EK, Moon HJ, Kwak JY, Kim MJ, Interobserver Variability of Ultrasound Elastography: How It Affects the Diagnosis of Breast Lesions. Am J Roentgenol 2011;196:730-36.

Zhang Q, Cai Y, Hua Y, Shi J, Wang Y, Wang Y, Sonoelastography shows that Achilles tendons with insertional tendinopathy are harder than asymptomatic tendons. Knee Surgery, Sports Traumatology, Arthroscopy 2016:DOI: 10.1007/s00167-016-4197-8.

Zhang Q, Li C, Han H, Yang L, Wang Y, Wang W, Computer-aided quantification of contrast agent spatial distribution within atherosclerotic plaque in contrast-enhanced ultrasound image sequences. Biomed Signal Process Control 2014;13:50-61.

Zhang Q, Li C, Han H, Yang L, Wang Y, Wang W, Computer-aided quantification of contrast agent spatial distribution within atherosclerotic plaque in contrast-enhanced ultrasound image sequences. Biomed Signal Process Control 2014;13:50-61.

Zhang Q, Xiao Y, Chen S, Wang CZ, Zhengy HR, Quantification of Elastic Heterogeneity Using Contourlet-Based Texture Analysis in Shear-Wave Elastography for Breast Tumor Classification. Ultrasound Med Biol 2015;41:588-600.

Zhang Q, Xiao Y, Dai W, Suo J, Wang C, Shi J, Zheng H, Deep learning based classification of breast tumors with shear-wave elastography. Ultrasonics 2016;72:150-57.

Zhang X, Xiao Y, Zeng J, Qiu W, Qian M, Wang C, Zheng R, Zheng H, Computer-assisted assessment of ultrasound real-time elastography: initial experience in 145 breast lesions. Eur J Radiol 2014;83:e1-7.

Zhao QL, Ruan LT, Zhang H, Yin YM, Duan SX, Diagnosis of solid breast lesions by elastography 5-point score and strain ratio method. Eur J Radiol 2012;81:3245-49.

Zhi H, Ou B, Xiao X-y, Peng Y-l, Wang Y, Liu L-s, Xiao Y, Liu S-j, Wu C-j, Jiang Y-x, Ultrasound elastography of breast lesions in chinese women: a multicenter study in China. Clinical breast cancer 2013;13:392-400.

460 Zhou J, Zhan W, Dong Y, Yang Z, Zhou C, Stiffness of the surrounding tissue of breast lesions evaluated by
461 ultrasound elastography. Eur Radiol 2014;24:1659-67.
462

464

Appendix

Feature Definitions

In shape features, the *convex area* is different from the tumor *area* and is defined as the area inside the convex polygon containing the tumor region. The equivalent diameter is equal to $\sqrt{4area/\pi}$, and the solidity is *area* / *convex area*.

In intensity statistics, the entropy of histogram (EtH) is given by

$$\mathrm{EtH} = -\sum_{i=0}^{255} p_i \log_2(p_i) \tag{A1}$$

Here p_i (i = 0, 1, …, 255) is the probability of intensity i in the image where the hue values have been requantized to 256 intensities. The area ratio (AR) and combined area ratio (CAR) are defined as

$$\mathrm{AR} = hard\ area / area \tag{A2}$$

$$\mathrm{CAR}=\mathrm{AR}\times \mathrm{DD}/\mathrm{CDD} \tag{A3}$$

where *hard area* is the area of the hard region within a tumor and calculated with adaptive thresholding of the hue values, DD is the dispersion degree, and CDD is the center deviation degree (Zhang et al. 2014).

The GLCM $G(i, j)$ represents the frequency of pairs of two pixels with intensities i and j (requantized to 8 gray levels), separated by a specific distance and direction. The GLCM is normalized to get the joint conditional probability density function $p(i, j) = G(i, j) / [\sum_i \sum_j G(i, j)]$, from which the GLCM texture features are derived:

$$\mathrm{Energy} = \sum_{i=1}^{8}\sum_{j=1}^{8} p(i, j)^2 \tag{A4}$$

$$\text{Contrast} = \sum_{i=1}^{8}\sum_{j=1}^{8}|i-j|^2\, p(i,j) \tag{A5}$$

$$\text{Homogeneity} = \sum_{i=1}^{8}\sum_{j=1}^{8}\frac{p(i,j)}{1+|i-j|} \tag{A6}$$

$$\text{Entropy} = -\sum_{i=1}^{8}\sum_{j=1}^{8} p(i,j)\log_2 p(i,j) \tag{A7}$$

In the contourlet texture feature extraction, a hue image is first decomposed with the contourlet transform into multiscale LP bands and multiscale multi-directional BP subbands. In this paper, two-scale decomposition is conducted, and the BP subbands at the first and second scales are derived along 8 and 4 directions, respectively. Instead from the original hue image, the intensity statistics and GLCM features are calculated from the LP band at the second scale and BP subbands at both scales, to serve as the contourlet texture features (Zhang et al. 2015).

Table 1. Typical features selected by the F_v-metric from seven clusters.

Cluster #	Feature Quantity	Typical Feature*	Benign	Malignant	*R*	*Fv*	*P*
1	65	EtH-DIR1	7.02±0.68	5.93±1.35	0.328	0.723	<0.0001
2	16	Eccentricity	0.81±0.10	0.73±0.18	0.637	0.419	0.001
3	3	CAR-LP	0.86±0.15	0.90±0.13	0.312	-0.207	0.138
4	47	Contrast-DIR2	0.77±0.30	0.44±0.23	0.670	0.880	<0.0001
5	36	EtH	7.31±0.35	6.63±0.66	0.709	0.894	<0.0001
6	194	Median-SRA1	0.0020±0.0003	0.0025±0.0002	0.821	-1.362	<0.0001
7	3	Solidity	0.98±0.02	0.95±0.05	0.442	0.422	0.001

*The number (1 or 2) after the names of features denotes the level of contourlet transform.

Table 2. The area under a receiver operating characteristic curve (AUC), classification sensitivity (SEN), specificity (SPC), accuracy (ACC) and Youden's index (YI) via the leave-one-out cross validation for computerized methods. The best results in the validation set are denoted in a bold font.

Methods	Validation					Training				
	AUC	SEN	SPC	ACC	YI	AUC	SEN	SPC	ACC	YI
All features	0.811	64.3	82.7	76.1	47.0	1.000±0.000	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
PCA-SVM	0.887	71.4	93.3	85.5	64.8	0.889±0.005	78.4±0.6	87.9±0.6	84.5±0.3	66.4±0.7
ManualSel	0.890	73.8	90.7	84.6	64.5	0.908±0.004	81.0±0.7	86.7±0.5	84.6±0.4	67.7±0.9
DPN-Hinge	0.870±0.005	74.7±2.2	**94.2±2.3**	87.2±1.0	68.9±1.5	0.941±0.006	81.8±1.8	97.6±0.7	91.9±0.8	79.4±1.9
DPN-Fisher	0.889	78.6	89.3	85.5	67.9	0.937±0.003	80.8±1.1	98.6±0.3	92.2±0.4	79.4±1.1
DPN-SVM	0.859	78.6	84.0	82.1	62.6	0.930±0.004	81.5±2.5	91.2±1.0	87.7±0.8	72.7±2.2
Cluster-R	0.885	83.3	89.3	87.2	72.7	0.927±0.005	88.1±0.9	88.1±0.6	88.1±0.5	76.2±1.1
Cluster-Fv	**0.917**	**85.7**	89.3	**88.0**	**75.0**	0.937±0.004	87.9±1.0	92.0±0.3	90.5±0.4	79.9±1.0
Cluster-P	0.897	83.3	89.3	87.2	72.7	0.926±0.003	85.8±0.6	93.3±0.2	90.6±0.2	79.1±0.6

Table 3. Classification performance on strain elastography in representative clinical publications.

Publications	Patient No.	Tumor No.	Method*	Sensitivity	Specificity	Accuracy	AUC
Cho et al. 2010	94	99	Q	95.0	75.0	78.8	0.879
Moon et al. 2010	140	140	Q	92.0	74.0	79.3	0.890
Zhao et al. 2012	155	187	Q	87.7	88.5	88.2	0.909
Alhabshi et al. 2013	168	168	Q	91.0	88.1	89.3	/
Zhi et al. 2013	1036	1150	G	86.4	80.8	83.5	0.860
Zhou et al. 2014	118	127	Q	38.2	93.1	69.3	0.669
Zhang et al. 2014	104	145	Q	92.5	94.9	93.8	0.960
Kim et al. 2015	100	109	G+B	72.7	98.0	95.4	0.916
Park et al. 2015	55	63	G	71.4	97.6	88.9	/
Fausto et al. 2015	120	129	Q	88.2	86.6	86.8	0.937
Hao et al. 2016	738	770	G+B	97.0	80.6	87.1	0.886
Redling et al. 2016	156	164	G+Q+B	95.0	85.0	88.8	/
Cluster-Fv	117	117	Q+S	85.7	89.3	88.0	0.917
Cluster-Fv without CV	117	117	Q+S	85.7	94.7	91.5	0.937

*Q: quantitative features on elastography; G: qualitative grading on elastography; B: Breast Imaging Reporting and Data System (BI-RADS) on conventional ultrasound; S: shape features on conventional ultrasound.
AUC: area under the receiver operating characteristic curve; CV: cross-validation.
Only Q results are listed here if a study reported both Q and G results. All studies except our study Cluster-Fv are performed without CV.

Fig. 1 An elastogram of a benign breast tumor, and the hardness retrieval and tumor segmentation on it. (a) Left: the composite elastogram, displayed as a translucent color image superimposed on a grayscale B-mode image; right: the same B-mode image. (b) The pure elastogram in color scales, calculated by subtracting the B-mode from the composite elastogram. (c) The tumor boundary detected in the B-mode. (d) and (e) The tumor boundary superimposed on the color and grayscale pure elastograms, respectively; the magenta areas in the latter denote the areas with invalid hardness values.

Fig. 2 A heat map depicting *Z*-scores of 364 radiomics features for 117 breast tumors, with cluster trees obtained from hierarchical clustering. The rows (features) are agglomerated into 7 clusters and the columns (samples) into 2 groups.

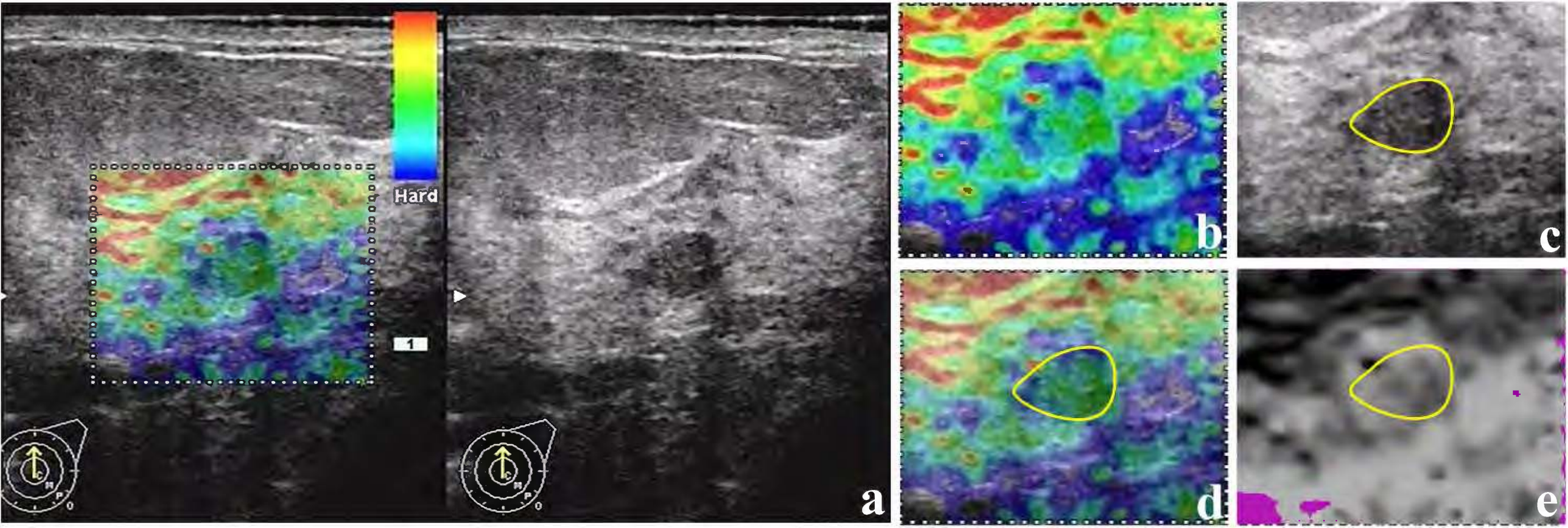
Fig. 3 Composite color elastograms (a, c) and grayscale B-mode images (b, d) of a malignant tumor (a, b) and a benign tumor (c, d).

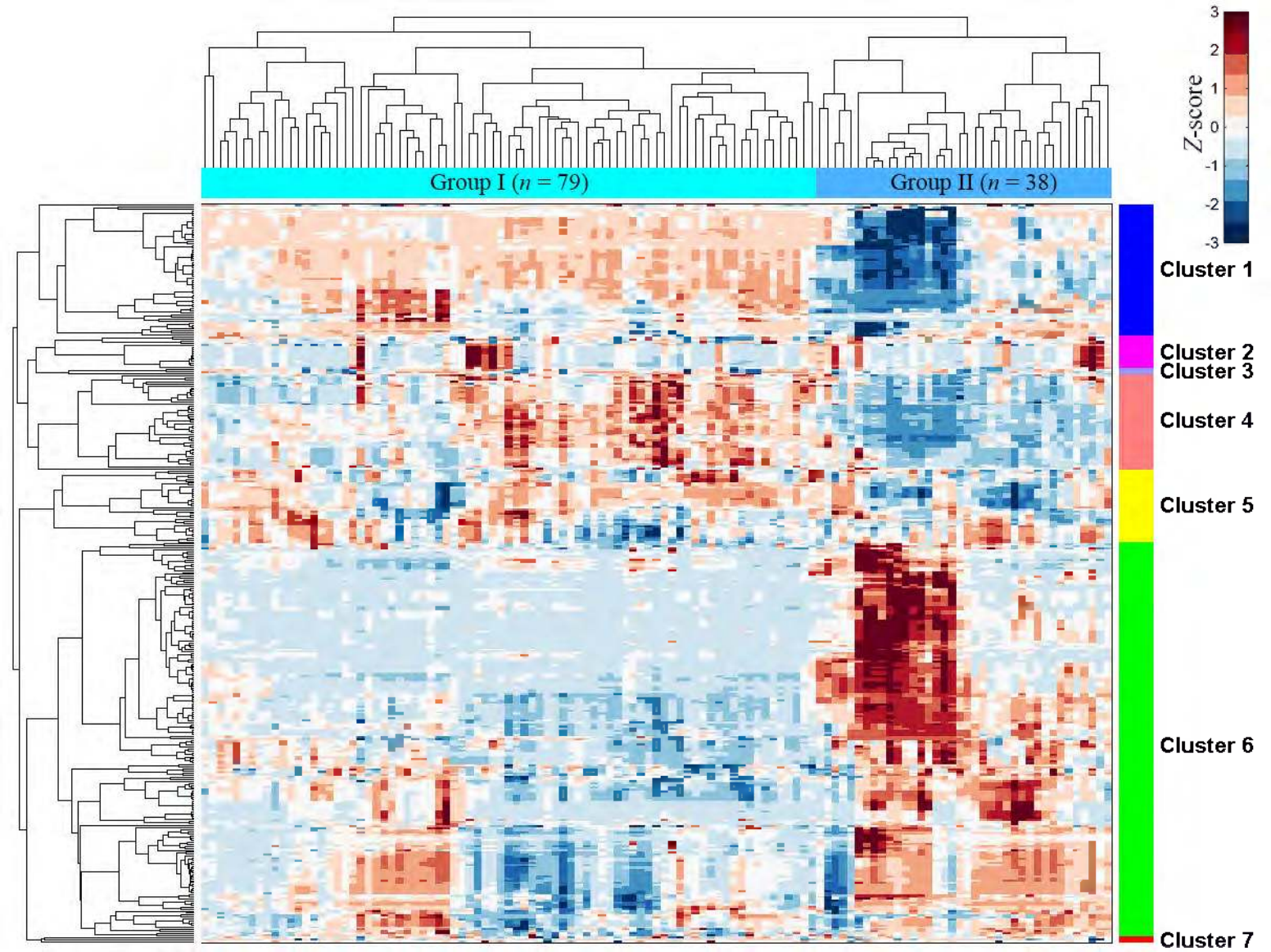
Fig. 4 The classification accuracy (ACC) and Youden's index (YI) of our classification scheme in the validation set when varying numbers of clusters from 2 to 15 with three feature selection metrics (R, F_v and P).

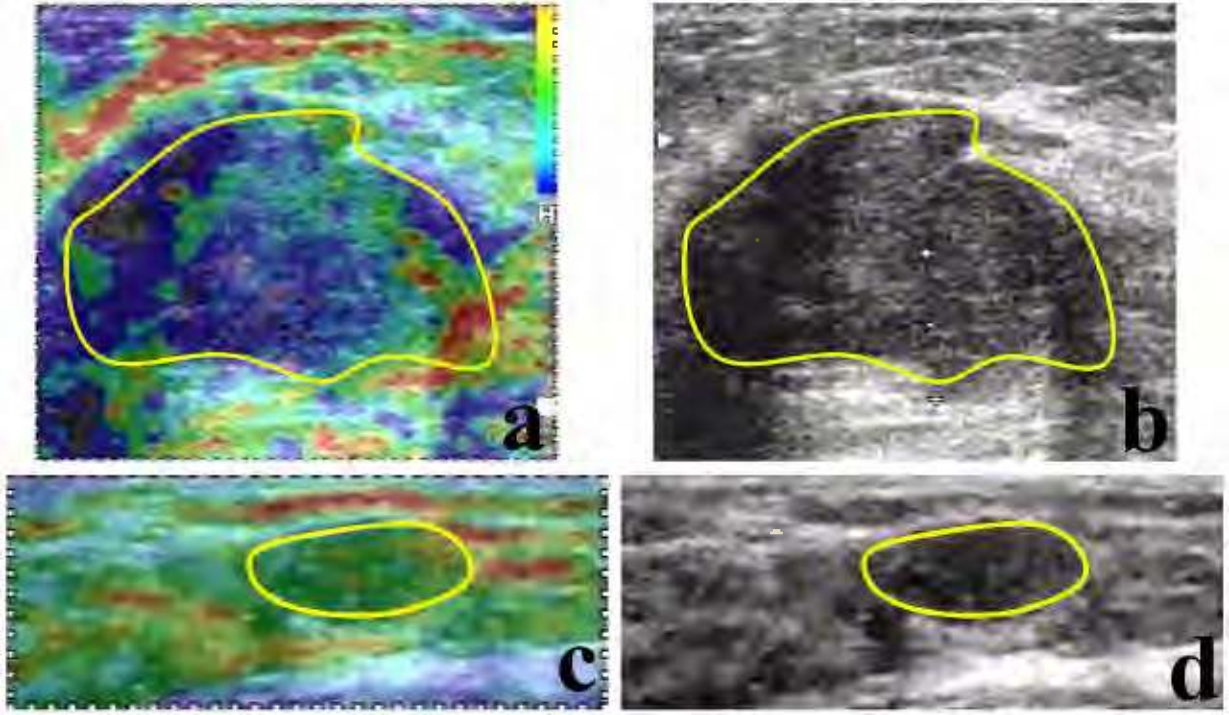
Fig. 5 Typical samples of malignant (a-j) and benign (k-t) breast tumors that were correctly classified with the proposed sonoelastomics method (Cluster-Fv).

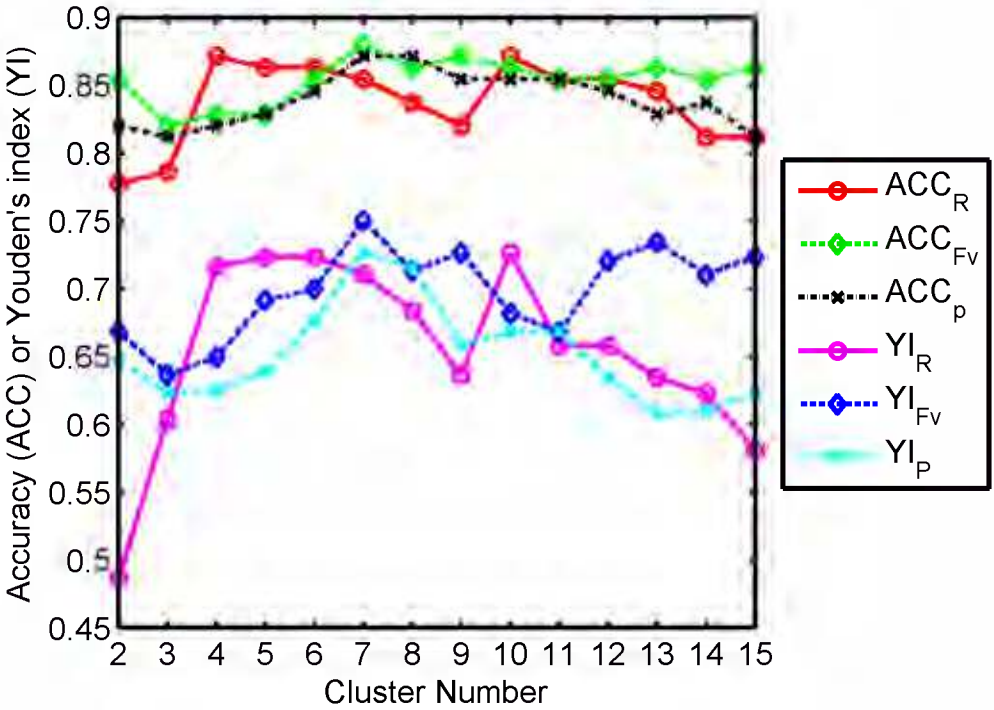
538

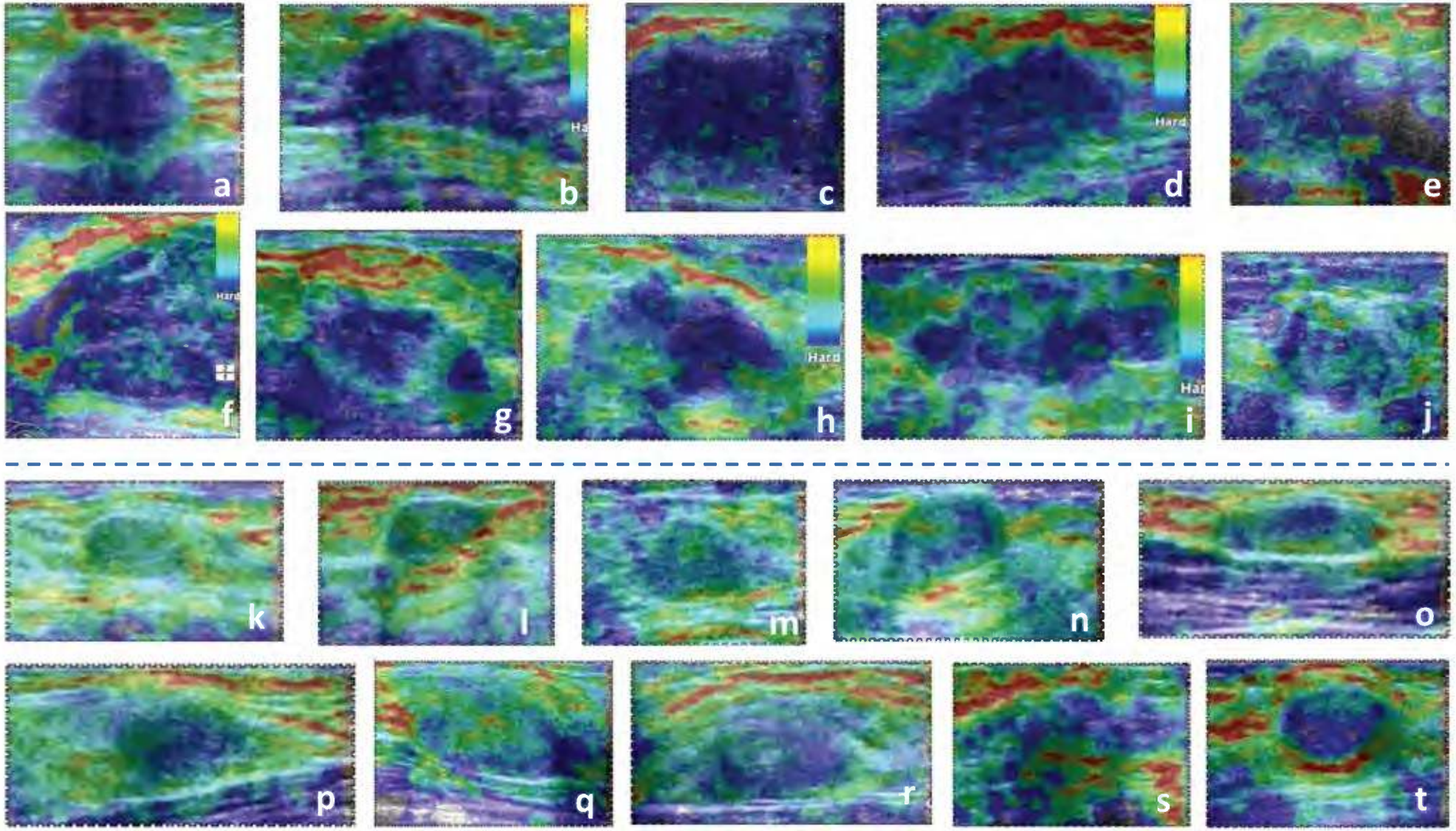
539 Fig. 6 The receiver operating characteristic curves of the proposed sonoelastomics method
540 (Cluster-Fv), the classic methods (All features, PCA-SVM and ManualSel), and the deep
541 learning method (DPN-Fisher).

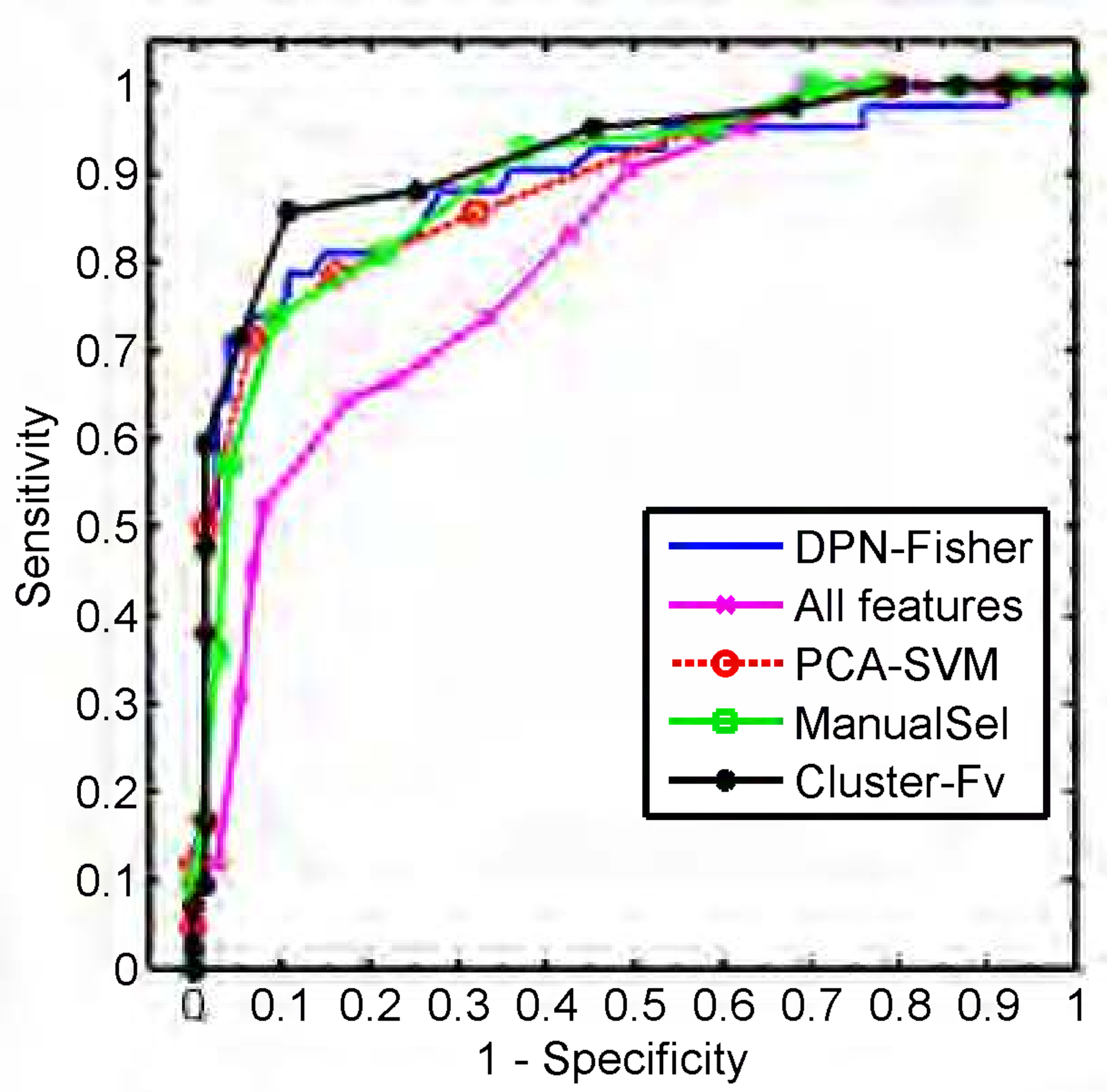

Hard
1
a
b
c
d
e


Group I (n = 79)
Group II (n = 38)
Z-score
3
2
1
0
-1
-2
-3
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7


a
b
c
d


0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
Accuracy (ACC) or Youden's index (YI)
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Cluster Number
ACC_R
ACC_Fv
ACC_p
YI_R
YI_Fv
YI_P


a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
q
r
s
t


Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 - Specificity
DPN-Fisher
All features
PCA-SVM
ManualSel
Cluster-Fv