Int. J. Appl. Math. Comput. Sci., 2020, Vol. 30, No. 2, 325–336
DOI: 10.34768/amcs-2020-0025

BOUNDED–ABSTAINING CLASSIFICATION FOR BREAST TUMORS IN IMBALANCED ULTRASOUND IMAGES

HONGJIAO GUAN[a,b,c], YINGTAO ZHANG[c,*], HENG-DA CHENG[d], XIANGLONG TANG[c]

[a] School of Cyber Security
Qilu University of Technology (Shandong Academy of Sciences)
3501 Daxue Road, Changqing District, Jinan 250353, China

[b] Shandong Key Laboratory of Computer Networks
3501 Daxue Road, Changqing District, Jinan 250353, China

[c] School of Computer Science and Technology
Harbin Institute of Technology
92 West Dazhi Street, Nangang District, Harbin 150001, China
e-mail: yingtao@hit.edu.cn

[d] School of Computer Science
Utah State University
Logan, UT 84322, USA

Computer-aided breast ultrasound (BUS) diagnosis remains a difficult task. One of the challenges is that imbalanced BUS datasets lead to poor performance, especially with regard to low accuracy in the minority (malignant tumor) class. Missed diagnosis of malignant tumors can cause serious consequences, such as delaying treatment and increasing the risk of death. Moreover, many diagnosis methods do not consider classification reliability; thus, some classifications may have a large uncertainty. To resolve such problems, a bounded-abstaining classification model is proposed. It maximizes the area under the ROC curve (AUC) under two abstention constraints. A total of 219 (92 malignant and 127 benign) BUS images are collected from the First Affiliated Hospital of Harbin Medical University, China. The experiment tests BUS datasets of three imbalance levels, and the performance contours are analyzed. The results demonstrate that AUC-rejection curves are less affected by class imbalance than accuracy-rejection curves. Compared with the state-of-the-art, the proposed method yields a significantly larger AUC and G-mean using imbalanced BUS datasets.

Keywords: breast ultrasound (BUS) images, reliable diagnosis, abstaining classification, imbalanced datasets.

1. Introduction

Breast cancer is the second leading cause of cancer death among women worldwide (Acharya *et al*., 2017; Monticciolo *et al*., 2017). Approximately 1 in 38 women die of breast cancer (https://www.cancer.org/cancer/breast-cancer). The pathology result of a biopsy is widely considered to be the golden standard for tumor diagnosis (Fu *et al*., 2018). However, biopsy is an invasive procedure that causes physical and mental stress to patients. Medical imaging techniques provide significant assistance toward early detection and diagnosis (Yassin *et al*., 2017). Ultrasonography is a popular detection means because it is non-radioactive, non-invasive, easily implemented, and low-cost (Rahmawaty *et al*., 2016).

Interpreting a large number of breast ultrasound (BUS) images is a cumbersome and repetitive task. Furthermore, the analysis of these images depends on the observer's personal skills and subjective experience, leading to inter- and intra-observer variations (Rawashdeh *et al*., 2018). Computer-aided diagnosis (CAD) techniques can improve the accuracy and objectivity of diagnosing breast tumors, and provide a second opinion to assist experts. In CAD systems, image enhancement

*Corresponding author

and speckle reduction are usually implemented to improve contrast and enhance edges. Histogram equalization and its variations are the main contrast enhancement methods (Mousania and Karimi, 2019; Shi *et al*., 2010). Wavelet transformation-based filtering shows the effectiveness of removing speckles and preserving the fine details of BUS images (Singh *et al.*, 2017a; Zhang *et al*., 2015). Image preprocessing is followed by feature representation and pattern classification. It has been demonstrated that morphological and texture features can effectively distinguish between benign and malignant breast tumors (Abdel-Nasser *et al*., 2017; Shan *et al*., 2016). Support vector machines (SVMs) (Cai *et al*., 2015; Daoud *et al*., 2016), neural networks (Lin *et al*., 2014; Shan *et al*., 2016), or AdaBoost (Zhou *et al*., 2013) are popular classification methods.

BUS diagnosis can be regarded as a binary classification problem. Although many studies on this type of diagnosis have shown promising classification performance, they ignore the important fact that the BUS characteristics of benign and malignant tumors have a significant overlap due to the heterogeneity in breast cancer (Chang *et al*., 2011; Chen *et al*., 2004; Garcia-Closas *et al.*, 2008; Li *et al*., 2017). In other words, some tumors are difficult to distinguish due to overlapping features. Therefore, forcing classification of ambiguous instances is unreasonable and can lead to errors. Furthermore, the cost of misclassifying breast tumors is very high. If the misclassification by the classifier is unavoidable, abstaining classification of uncertain and difficult examples can ensure reliable outputs. The classified examples are assigned their class labels with a large certainty. The rejected examples could have ambiguous features of benign or malignant tumors; therefore, such tumors should be studied more closely. Performing other examinations and organizing expert consultation will likely provide the appropriate outcome.

Imbalanced datasets, which are a common occurrence in the field of medical diagnosis, pose another problem. The number of malignant (positive class) tumors is commonly smaller than that of benign (negative class) ones. Some abstaining classification methods minimize the average cost, and thus they are required to be equipped with information on costs. However, in practice, it is difficult to obtain or estimate misclassification costs. Some abstaining classification methods aim to obtain high accuracy or a low error rate for a given reject rate (Fischer *et al*., 2015; Kang *et al*., 2017; Pietraszek, 2007; Wang *et al*., 2017). However, when datasets are imbalanced, obtaining high accuracy may not be relevant. The optimization metrics (accuracy or the error rate) can guide learning algorithms to be biased toward the majority (negative) class (López *et al*., 2013). Consequently, the recognition rate of the minority (positive) class is low. When the negative class size is much larger than that of the positive class, if all examples are classified as negative in an extreme case, the overall accuracy may be high. However, in this case, all the positive examples will be misclassified. That is, all malignant tumors will be missed. Imbalanced datasets are prevalent not only in medical diagnosis (Yu and Ni, 2014), but also in other safety-critical fields such as intrusion detection (Tesfahun and Bhaskari, 2013).

To address such problems, we propose an abstaining classification model with two abstention constraints. The conditional model maximizes the AUC metric instead of accuracy. This allows the model to adapt to situations where datasets are imbalanced. Note that the AUC denotes the area under the receiver operating characteristic (ROC) curve. Furthermore, the proposed method avoids setting costs which are usually unknown in practice. The performance-abstention trade-off curves are analyzed through experiments using different imbalance level BUS datasets.

2. Methods

Figure 1 shows the entire process of BUS diagnosis discussed in this paper. In Section 2.1, the first two steps, i.e., denoising to reduce speckle and annotating tumor boundaries, are introduced. Section 2.2 introduces the morphological and texture features used in the experiments. Section 2.3 describes the infinite latent feature selection (ILFS) (Roffo *et al*., 2017) algorithm in brief. The proposed method and related works are presented in Section 2.4. The results of using the leave-one-out cross-validation (LOOCV) procedure are reported and analyzed in Section 3.

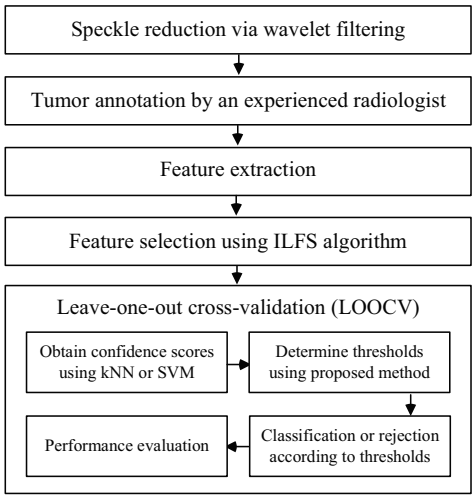


Fig. 1. BUS diagnosis procedure in this paper.

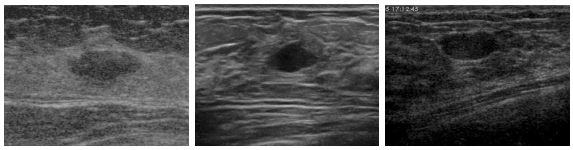
2.1. Image acquisition and preprocessing. A BUS image database including 219 (92 malignant and 127 benign) images was collected from the First Affiliated Hospital of Harbin Medical University, China, using a VIVID 7 ultrasound scanner (GE, Horten, Norway) with a linear probe of 5–14 MHz. The tumor type (benign or malignant) was confirmed through surgery and pathological examination. Informed consent was obtained from each participant and the patient's privacy was well protected. The study was approved by the related Institutional Review Board.

Ultrasound images suffer from speckle noise, which degrades image quality and affects post-processing steps, including feature extraction and tumor classification (Cheng *et al*., 2010; Gai *et al*., 2018; Roy *et al*., 2018; Singh *et al.*, 2017b). Haar wavelet-based wavelet filtering was used to reduce speckles and preserve details of tumor images (Singh *et al.*, 2017a). The two-level wavelet decomposition was performed for each BUS image. Consequently, approximation coefficients (LL2) and detail coefficients (HH1, HL1, LH1, HH2, HL2, and LH2) were generated. High frequency and speckle noise are commonly contained in HH bands. Here, the HH1 and HH2 coefficients were eliminated, and the residual coefficients were used to reconstruct the denoised image.

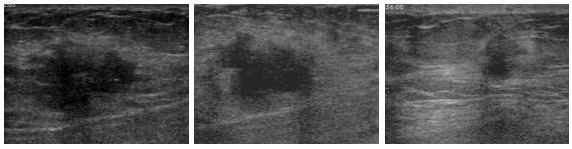
Segmentation of tumor regions is a difficult task due to various shapes, sizes, and positions of lesions. In this study, the tumor boundaries of the BUS images were manually delineated by an experienced radiologist from the First Affiliated Hospital of Harbin Medical University, China. Figure 2 shows samples of benign and malignant tumors, respectively.

2.2. Feature extraction. The Breast Imaging Reporting and Data System (BI-RADS) provides a standard characterization of breast tumor images (Liberman and Menell, 2002). The BI-RADS describes breast tumors based on shape, boundary, orientation, echo pattern, acoustic shadowing, and so on (Rodriguez-Cristerna *et al*., 2018; Yu *et al*., 2018). According to the system, benign tumors are usually round or elliptical, with smooth and clear borders and homogeneous internal echoes. Malignant tumors typically appear as irregular shapes, blurry borders, spiculated margins, and heterogeneous internal echoes with posterior acoustic shadowing.

Morphology and texture features are two main categories that describe and analyze breast tumor images (Lee *et al*., 2018). According to the BI-RADS, 13 morphological features describing the shape, margin, and orientation of tumors were extracted from each BUS image (Cheng *et al*., 2010; Moon *et al*., 2018). In addition, Haralick's texture features were extracted (Haralick *et al*., 1973). Table 1 summarizes these features. A total of 168 Haralick's texture features $t_1 - t_{168}$ were computed



(a) benign tumor images



(b) malignant tumor images

Fig. 2. Samples of tumor images.

based on the gray-level co-occurrence matrix (GLCM) (Haralick *et al*., 1973). Each GLCM element $\boldsymbol{P}_{d,\theta}(i,j)$ is the joint probability of the tumor gray values at distance d in direction θ. Specifically, the GLCM is computed as follows (Liu *et al*., 2012):

$$\begin{aligned}\boldsymbol{P}_{d,\theta}(i,j) = \|\{(x_1,y_1),(x_2,y_2)\}| \\ \{x_2 - x_1 = d\cos\theta, y_2 - y_1 = d\sin\theta, \\ \boldsymbol{I}(x_1,y_1) = i, \boldsymbol{I}(x_2,y_2) = j\}\|, \quad (1)\end{aligned}$$

where (x_1,y_1) and (x_2,y_2) are the positions of two pixels in the tumor, $\boldsymbol{I}$ is the gray matrix of the tumor, and $\|\cdot\|$ denotes the number of pixel pairs satisfying the conditions in Eqn. (1). Based on the GLCM, fourteen feature descriptors were calculated: angular second moment, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measure of correlation (I, II), maximal correlation coefficient. In this study, three distances (d =1, 2, 3) and four directions ($\theta = 0°, 45°, 90°, 135°$) were employed; therefore, 168 (3×4×14) GLCM-based Haralick's features were obtained. In addition to the 13 morphological features, a total of 181 features were extracted from each BUS tumor image.

2.3. Feature selection. Too many features can increase model complexity and computational cost. Furthermore, using more features has a greater risk of overfitting. Feature selection can improve classification performance by removing irrelevant or redundant features. Feature selection methods can be mainly divided into three categories: wrappers, embedded methods, and filters (Guyon and Elisseeff, 2003). Wrapper feature selection methods are dependent on specific classification algorithms and utilize the classification performance to evaluate the candidate feature subsets. Embedded methods perform feature selection in the process of model training. Filter feature selection methods contain two steps: ranking and selecting. In the first step, all

Table 1. Summary of morphological and texture features.

Morph. features	Description
Shape	
s_1	circularity
s_2	length-to-width ratio of enclosing rectangle
s_3	ratio of max-to-min radial length
s_4	average of normalized radial length
s_5	standard deviation of norm. radial length
s_6	entropy of normalized radial length
s_7	area ratio
Boundary	
b_1	boundary roughness
b_2	number of lobulations
b_3	spiculation
b_4	elliptic-normalized circumference
b_5	elliptic-normalized skeleton
Orientation	
θ	angle
Texture features	**Description**
$t_1 - t_{12}$	angular second moment (energy)
$t_{13} - t_{24}$	contrast
$t_{25} - t_{36}$	correlation
$t_{37} - t_{48}$	variance
$t_{49} - t_{60}$	inverse difference moment (homogeneity)
$t_{61} - t_{72}$	sum average
$t_{73} - t_{84}$	sum variance
$t_{85} - t_{96}$	sum entropy
$t_{97} - t_{108}$	entropy
$t_{109} - t_{120}$	difference variance
$t_{121} - t_{132}$	difference entropy
$t_{133} - t_{144}$	information measure of correlation I
$t_{145} - t_{156}$	information measure of correlation II
$t_{157} - t_{168}$	maximal correlation coefficient

m candidate features are ranked according to a score assignment strategy. Then, the top $\tilde{m}$ ($\tilde{m} \ll m$) features are chosen using cross-validation. Filter methods are usually faster than wrapper methods, and they are independent of specific algorithms. Therefore, in the study, a filter method was chosen for feature selection.

The infinite latent feature selection (ILFS) algorithm is a popular filter method. To rank all features, three steps were enforced: preprocessing, graph-weighting, and ranking. During the preprocessing, each feature $\vec{x_i}$ was mapped to a descriptor $\vec{f_i}$ using a discriminative quantization (DQ) process. Assume that there were n examples and m features. Both $\vec{x_i}$ and $\vec{f_i}$ were $n \times 1$ vectors. The number of distinct values in $\vec{x_i}$ may be very large. Thus, the DQ process mapped the large set of raw values as a smaller set that contains countable tokens. In the second step, an undirected fully connected graph was built. In the graph, each node denoted a feature descriptor f_i, and the weighted edge between f_i and f_j represented the probability that features x_i and x_j were relevant. The weights were obtained by learning a probabilistic latent semantic analysis (PLSA) model (Hofmann, 1999). The PLSA model introduced two topics, which represented two latent variables: relevancy and irrelevancy. Thus, PLSA built a simple Bayesian network and modeled the probability of token-descriptor co-occurrences. After deriving the weights between every two nodes, the joint probability of each path of length l ($l = 1, 2, \dots$) was estimated, and finally, features were ranked according to descending scores, following the idea of the infinite feature selection algorithm (Roffo *et al.*, 2015). For details of the ILFS algorithm, please refer to the work of Roffo *et al.* (2017).

2.4. Abstaining classification.

2.4.1. Related works. In the traditional binary classification problem, the classification rule is as follows:

$$\text{class} = \begin{cases} \text{positive} & \text{if } s(m) > t, \\ \text{negative} & \text{otherwise,} \end{cases} \tag{2}$$

where t is the decision threshold and $s(m)$ is the confidence score of example m belonging to the positive class. In the classification with reject option, the classification rule is as follows:

$$\text{class} = \begin{cases} \text{positive} & \text{if } s(m) > t_2, \\ \text{negative} & \text{if } s(m) \le t_1, \\ \text{reject} & \text{otherwise.} \end{cases} \tag{3}$$

Confidence scores can be obtained using traditional classifiers (Fawcett, 2004; 2006). In this paper, the k-nearest neighbor (kNN) and SVM algorithms are used as scoring classifiers to generate confidence scores.

Abstaining classification models are built to determine the rejection thresholds t_1 and t_2 ($t_1 < t_2$). Tortorell (2000; 2004) proposes an abstention model, which minimizes the average cost:

$$\min_{t_1,t_2} \ \text{cost}(t_1, t_2), \tag{4}$$

where

$$\begin{aligned} &\text{cost}(t_1, t_2) \\ &= p(+) \cdot \text{CFN} \cdot \text{fnr}(t_1) + p(-) \cdot \text{CTN} \cdot \text{tnr}(t_1) \\ &\quad + p(+) \cdot \text{CTP} \cdot \text{tpr}(t_2) + p(-) \cdot \text{CFP} \cdot \text{fpr}(t_2) \\ &\quad + p(+) \cdot \text{CRP} \cdot \text{rpr}(t_1, t_2) \\ &\quad + p(-) \cdot \text{CRN} \cdot \text{rnr}(t_1, t_2). \end{aligned} \tag{5}$$

Here $p(+)$ and $p(-)$ are the prior probabilities of the positive and negative classes, respectively, CTN, CTP, CFP, and CFN are the costs of true negatives, true positives, false positives, and false negatives, respectively, 'tnr' and 'fpr' are the true negative rate and false positive rate among all negative examples, respectively, 'tpr' and 'fnr' denote the true positive rate and false negative rate among all positive examples, respectively, CRP and CRN indicate the rejection costs regarding positive and negative classes, respectively, and 'rpr' and 'rnr' are the ratios of rejected positive and negative examples, respectively. In such abstention models, the costs of classification and rejection are required to be known. However, costs are unbounded values, which are usually difficult to obtain or estimate in practice.

Pietraszek (2007) proposes a bounded-abstention (BA) model, which adds a rejection constraint. The BA model is as follows:

$$\begin{aligned} &\min_{t_1,t_2} \ \frac{\text{CFN} \cdot FN(t_1) + \text{CFP} \cdot FP(t_2)}{TN(t_1) + FP(t_2) + TP(t_2) + FN(t_1)} \\ &\text{subject to} \\ &\qquad \text{rej}(t_1, t_2) \le k_{\max}, \end{aligned} \tag{6}$$

where FN (FP) means the number of false negatives (positives), TN (TP) is the number of true negatives (positives), 'rej' denotes the overall reject rate, i.e., the number of rejected examples divided by the sample size. When misclassification costs CFN and CFP are the same, the BA model in Eqn. (6) starts minimizing the error rate under a reject rate constraint. In such a case, the BA model is not suitable to deal with imbalanced datasets since the optimization of the error rate can lead to a biased classification performance.

2.4.2. Proposed abstention model. The ROC/AUC is insensitive to an imbalanced class distribution and unequal misclassification costs (Fawcett, 2004; Prati *et al*., 2011). The proposed model aims to obtain the maximum AUC under two abstention constraints (Guan *et al*., 2019). The optimization problem is formalized as follows:

$$\begin{aligned} &\max_{t_1,t_2} \ \text{AUC}(t_1, t_2) \\ &\text{subject to} \\ &\qquad \begin{cases} \text{rnr}(t_1, t_2) \le n_{\max}, \\ \text{rpr}(t_1, t_2) \le p_{\max}. \end{cases} \end{aligned} \tag{7}$$

$\text{AUC}(t_1, t_2)$ is defined as (Hong *et al*., 2007; López *et al*., 2013)

$$\begin{aligned} \text{AUC}(t_1, t_2) &= \frac{1 + \text{tpr}(t_1, t_2) - \text{fpr}(t_1, t_2)}{2} \\ &= \frac{\text{tpr}(t_1, t_2) + \text{tnr}(t_1, t_2)}{2}. \end{aligned} \tag{8}$$

Note that 'tpr' in Eqn. (8) is different from that in Eqn. (5). Here, 'tpr' is the ratio of true positives in the classified positive examples. Similarly, 'fpr' and 'tnr' in Eqn. (8) are the ratios of false positives and true negatives among the classified negative instances, respectively. The AUC distinguishes the accuracies of the positive and negative classes. Thus, the AUC can evaluate the classifier performance impartially when confronting imbalanced datasets. Additionally, $n_{\max}$ and $p_{\max}$ are the hyperparameters of the proposed model. The values of $n_{\max}$ and $p_{\max}$ are in the interval of [0,1], so the hyperparameters are bounded.

A method of solving the proposed optimization problem is the exhaustive algorithm; that is, to try every possible combination of t_1 and t_2. Assume that $s_{\min}$ and $s_{\max}$ are the minimum and maximum values of the scores of all training examples, respectively. The exhaustive algorithm assigns t_1 or t_2 scores between $s_{\min}$ and $s_{\max}$ with step $(s_{\max} - s_{\min})/k$ and computes the corresponding $\text{AUC}(t_1, t_2)$, $\text{rpr}(t_1, t_2)$, and $\text{rnr}(t_1, t_2)$. Finally, the combination of t_1 and t_2 that has the largest AUC and simultaneously satisfies the two constraints is the empirical solution to the optimization problem. In the experiment, k is set as 200.

3. Results and a discussion

3.1. Leave-one-out cross-validation. The size of the original BUS dataset (a total of 219 examples) is not very large. Moreover, in Sections 3.4 and 3.5, BUS datasets of different imbalance levels will be adopted. The sample size will become smaller. Therefore, in the following experiments, the leave-one-out cross-validation procedure is performed. Each time an example was selected as the test example, and the remaining ones were used as training examples to determine two rejection thresholds. In the test phase, the score of the test example was first generated using the scoring classifier. Then, the test example was classified or rejected according to Eqn. (3). The entire process was executed n times until each example ran as the test example; n is the sample size of the dataset.

3.2. Evaluation metrics. In the paper, the classification performance was evaluated using accuracy (ACC), AUC, G-mean, sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), Matthew's correlation coefficient (MCC). Their definitions are as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP+FN+TN+FP}}, \tag{9}$$

$$\mathrm{AUC} = \frac{\mathrm{SEN}+\mathrm{SPE}}{2}, \tag{10}$$

$$\text{G-mean} = \sqrt{\mathrm{SEN}\times\mathrm{SPE}}, \tag{11}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{12}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \tag{13}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{14}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}, \tag{15}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{16}$$

where TP and TN are the numbers of positive and negative examples that are correctly classified, respectively, while FN and FP are the numbers of positive and negative examples that are misclassified, respectively.

3.3. Selecting features. Both the kNN and SVM algorithms were used to classify the BUS dataset (219 examples) in the LOOCV procedure. After testing multiple parameters, the experimental setups of the two scoring classifiers were as follows. The nearest neighbor size $k = 5$ was employed in kNN, and the linear kernel function was used in the SVM. The ILFS algorithm ranked 181 features, and finally the top 19 features were selected since using 19 features yielding better results than employing 181 features. The results of the kNN and SVM algorithms are summarized in Table 2. In the following experiments (Sections 3.4 and 3.5), the selected 19 features were employed. They are t_{54} (inverse difference moment, $d = 2, \theta = 45°$), t_{161} (maximal correlation coefficient, $d = 2, \theta = 0°$), t_{164} (maximal correlation coefficient, $d = 2, \theta = 135°$), t_{57} (inverse difference moment, $d = 3, \theta = 0°$), t_{98} (entropy, $d = 1, \theta = 45°$), t_{100} (entropy, $d = 1, \theta = 135°$), t_{103} (entropy, $d = 2, \theta = 90°$), t_{58} (inverse difference moment, $d = 3, \theta = 45°$), b_5 (elliptic-normalized skeleton), b_3 (spiculation), b_2 (number of lobulations), s_1 (circularity), θ (angle), t_{146} (information measure of correlation II, $d = 1, \theta = 45°$), s_2 (length-to-width ratio of enclosing rectangle), t_{148} (information measure of correlation II, $d = 1, \theta = 135°$), t_{149} (information measure of correlation II, $d = 2, \theta = 0°$), t_{147} (information measure of correlation II, $d = 1, \theta = 90°$), and t_{145} (information measure of correlation II, $d = 1, \theta = 0°$).

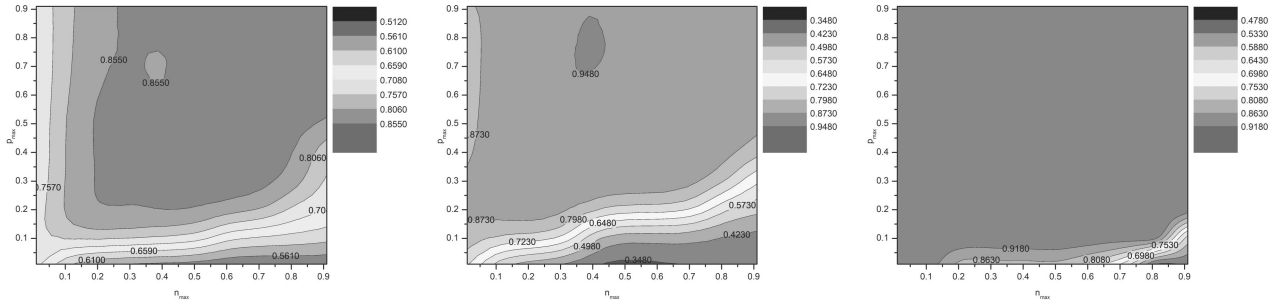
Note that, as the top 19 features were chosen using all 219 examples, these features could not be the best ones for the imbalanced datasets in Sections 3.4 and 3.5. However, considering the small sample size, the number of futures (i.e., 181) was rather large. Furthermore, we stress that the following experiments aim to demonstrate the advantage of the proposed method and not to discuss the best features. Therefore, in the following experiments, the chosen 19 features are used.

3.4. Results of using different imbalance level BUS datasets. In this experiment, we aim to explore the performance-abstention trade-off when using different imbalance level datasets, and to discuss how the rejection parameters $p_{\max}$ and $n_{\max}$ may be chosen. Here, BUS datasets at three imbalance levels were used. Among 219 BUS examples, 64 (25 or 13) positive examples were randomly drawn and combined with all 127 negative examples to form a 1:2 (1:5 or 1:10) imbalanced BUS dataset. For the proposed method, the parameters $p_{\max}$ and $n_{\max}$ were set from 0.01 to 1 with step 0.1. The accuracy and the AUC were used as performance metrics. The imbalanced datasets were randomly drawn ten times, and the obtained metrics were averaged. Finally, iso-performance profiles were plotted to display the performance change with varying $p_{\max}$ and $n_{\max}$. Two scoring classifiers, kNN and the SVM, were used.

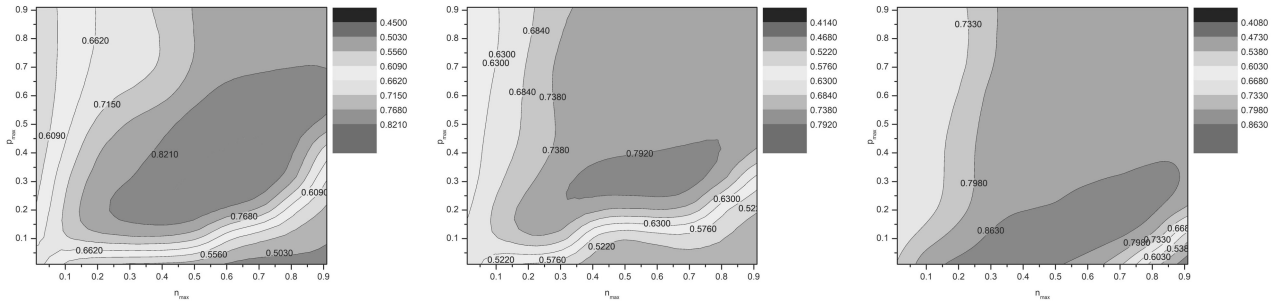
Figure 3 shows the accuracy and AUC contours of using kNN to classify the 1:2, 1:5, and 1:10 BUS datasets (corresponding to the figures from left to right in each row). In the subgraphs, the horizontal and vertical axes denote the parameters $n_{\max}$ and $p_{\max}$, respectively.

Table 2. Results of kNN and the SVM using the 19 selected features and all morphological and texture features.

	ACC	AUC	G-mean	SEN	SPE	PPV	NPV	MCC
KNN								
selected(19)	**82.65%**	**81.59%**	**81.33%**	**75.00%**	**88.19%**	**82.14%**	**82.96%**	**64.14%**
all(181)	73.52%	71.47%	70.32%	58.70%	84.25%	72.97%	73.79%	44.82%
SVM								
selected(19)	**87.67%**	**87.12%**	**87.06%**	**83.70%**	**90.55%**	**86.52%**	**88.46%**	**74.61%**
all(181)	87.21%	86.58%	86.49%	82.61%	90.55%	86.36%	87.79%	73.65%

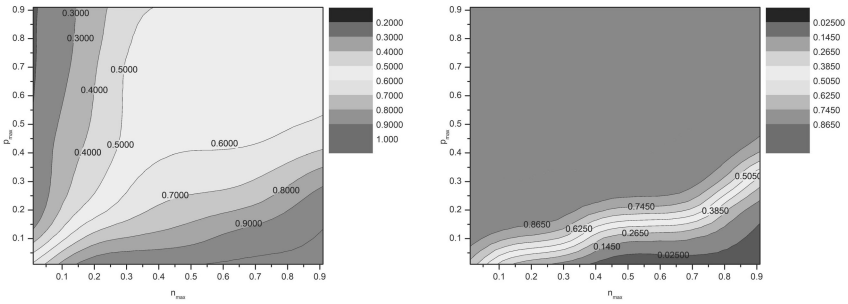


(a) accuracy contours



(b) AUC contours

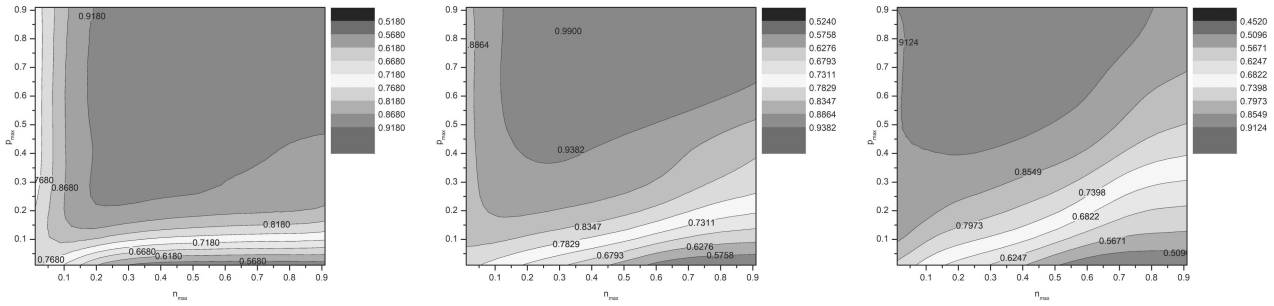
Fig. 3. iso-Accuracy and iso-AUC contours on 1:2, 1:5, and 1:10 BUS datasets (from left to right in each row) using kNN.



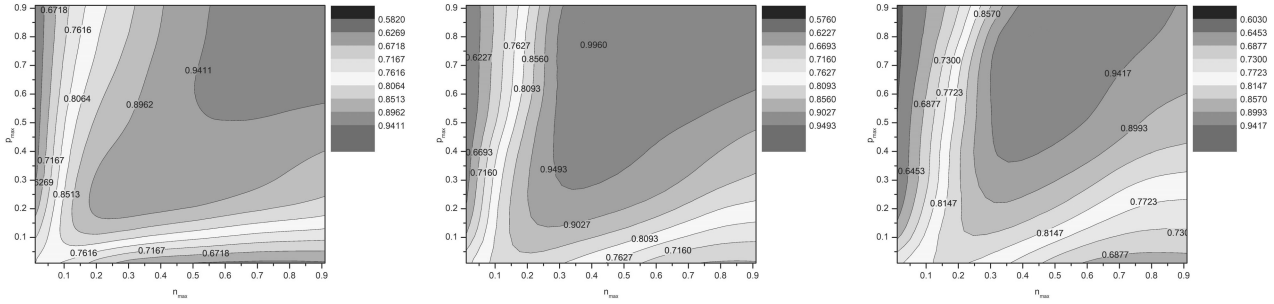
(a) sensitivity contours (b) specificity contours

Fig. 4. iso-SEN and iso-SPE ontours of 1:5 imbalance level using kNN.

Obviously, the larger the imbalance level, the denser the iso-accuracy lines at the bottom of the graph. For the AUC contours, such a phenomenon is less serious. This indicates that AUC is less affected by the class distribution than accuracy. This may be explained using the sensitivity (SEN) and specificity (SPE) contours. Figure 4 only displays the iso-SEN and iso-SPE curves of the imbalance level 1:5. The iso-SPE contours are located near the horizontal axis. For a fixed $n_{\max}$, a small change in $p_{\max}$ causes a large change in the SPE values. In contrast,



(a) accuracy contours



(b) AUC contours

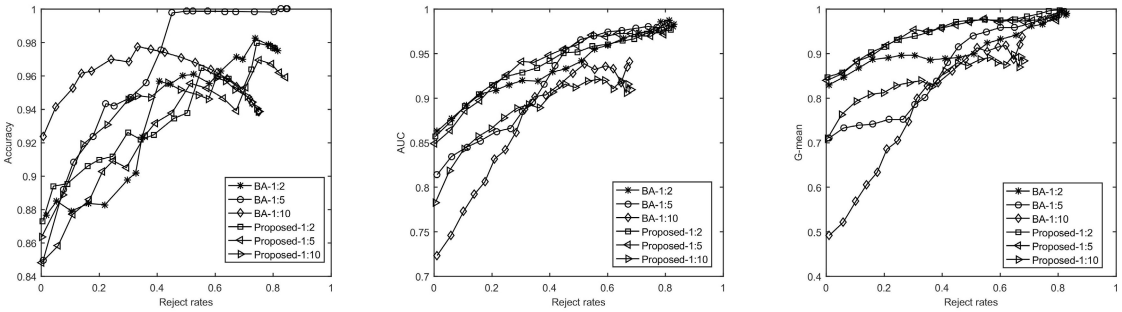
Fig. 5. iso-Accuracy and iso-AUC contours on 1:2, 1:5, and 1:10 BUS datasets using the SVM.

when $p_{\max}$ is fixed, the change in SPE values is small as $n_{\max}$ increases. This implies that SPE is more sensitive to $p_{\max}$ than to $n_{\max}$. The iso-SEN contours are less biased than the iso-SPE ones. According to the definitions of accuracy and the AUC, the bias has a greater impact on accuracy than on the AUC since the negative class dominates in sample size. In addition, when $p_{\max}$ is fixed and $n_{\max}$ increases, SEN increases and SPE decreases; when $n_{\max}$ is fixed and $p_{\max}$ increases, SPE increases and SEN decreases. However, the changes in the magnitudes of SEN and SPE are not consistent. Hence, when both $p_{\max}$ and $n_{\max}$ increase, the gradient direction of the iso-AUC contours does not follow the $y = x$ line. In cost-sensitive applications, where the positive class has higher error cost than the negative class, one can set $n_{\max}$ larger than $p_{\max}$ to obtain high sensitivity.

Figure 5 displays the accuracy and AUC contours using the SVM as the scoring classifier. The contours are similar to those obtained when using kNN. However, one difference is noted: the gradient directions of the iso-AUC contours remain almost unchanged when the imbalance level becomes large. Furthermore, the direction roughly follows $y = x$. This is because the SVM is less susceptible to class imbalance than kNN (Wu and Chang, 2005). An important point is that using the SVM yields better classification performance than using kNN.

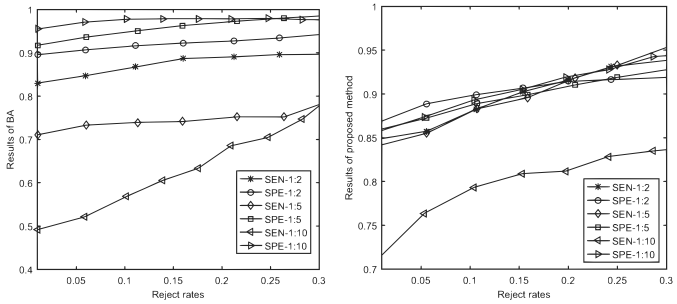
To summarize, the performance of the proposed abstaining classifier is associated with scoring classifiers, datasets to be classified, and abstention parameters. To obtain good results, one should choose the appropriate scoring classifier. For a specific application (such as BUS diagnosis), the data distribution can basically be determined when a larger number of cases are collected impartially. For example, we can collect BUS images of patients who undergo breast cancer screening (not necessarily malignant) and treat the disease (malignant). Once the data and the scoring classifier are determined, the performance-abstention curve is obtained. Therefore, the practitioner can choose the abstention parameters according to the trade-off curve. If he/she focuses on the classification performance, it is possible to select a larger abstention. If the resources for addressing the rejected cases are limited, one can use a small abstention.

3.5. Comparison of the proposed method with the state-of-the-art. The proposed method is compared with Pietraszek's BA model (Pietraszek, 2007). BA determines rejection thresholds by minimizing the misclassification cost sum of false positives and false negatives, and constraining the overall reject rate. If the misclassification costs of the two classes are considered identical, that is, the cost ratio of false positives to false negatives is 1, then the optimization goal of BA becomes the error rate. To ensure the comparability of BA and



(a) accuracy (b) AUC (c) G-mean

Fig. 6. Performance-rejection trade-off curves of BA and the proposed method.



(a) results of BA (b) results of the proposed method

Fig. 7. Results of sensitivity (SEN) and specificity (SPE) of BA and the proposed method when reject rates are lower than 0.3.

the proposed method, the same $p_{\max}$ and $n_{\max}$ were set from 0.01 to 1 with step 0.05, and the cost ratio in BA was set at 1. The overall reject rate in BA was also set from 0.01 to 1 with step 0.05. The comparative experiment was implemented using 1:2, 1:5, and 1:10 BUS datasets. The SVM was chosen as the scoring classifier because it provided better results than kNN. The performance-abstention curves of the two abstaining classifiers are shown in Fig. 6, where the horizontal axis denotes the resulting reject rate and the vertical axis represents the results of accuracy (a), AUC (b), and G-mean (c), respectively.

In Fig. 6(a), it is observed that BA yields better accuracy than the proposed method. The more unbalanced the class distribution, the larger the performance difference between the two methods. This result is attributed to the effect of class imbalance on the optimization target. Given a fixed reject rate, BA maximizes the accuracy, whereas the proposed method maximizes the AUC. When the imbalance level becomes larger, the influence of the minority class on the accuracy is weakened to a greater extent than on the AUC. This can be explained in Fig. 7, which shows the results of sensitivity and specificity when reject rates are lower than 0.3. In Fig. 7(a), the sensitivity values of BA change considerably with increasing imbalance levels. When the imbalance level is 1:10, the sensitivity values at abstention lower than 0.1 are less than 0.6. However, all the specificity values are higher than 0.9. Hence, the results of sensitivity and specificity are very imbalanced in BA. For the proposed method in Fig. 7(b), the values of sensitivity and specificity do not display as large differences as for BA. Therefore, the proposed method achieves better AUC-abstention (Fig. 6(b)) and G-mean-abstention (Fig. 6(c)) trade-offs than BA; for a fixed reject rate, the proposed method has a higher AUC or G-mean value than BA when reject rates are less than 0.3. Note that choosing low reject rates is usually significant in practical applications (Simeone *et al.*, 2012).

The areas under the performance (ACC, AUC, and G-mean)-abstention curves in the abstention range of [0,0.3] were estimated, and the mean and standard deviation were calculated (Table 3). The larger the average value, the better the classification performance. In addition, the Wilcoxon signed rank test was performed to compare BA and the proposed method, and the p-values are listed in Table 3. The results shown in bold are statistically significantly better. The results in Table 3 are consistent with those in Fig. 6. The proposed method performs significantly better than BA in terms of the AUC and G-mean when the reject rate is in the range of [0,0.3].

4. Conclusions

In BUS diagnosis, reliable classification is essential because of the high costs of missed diagnosis

Table 3. Results of mean and standard deviation of areas under performance-rejection curves and the p-values associated with the Wilcoxon signed rank test.

Area under ACC-abstention curves			
	BA	Proposed	p-Value
1:2	0.2581 (0.002)	**0.2689 (0.001)**	0.002
1:5	**0.2707 (0.003)**	0.2649 (0.002)	0.002
1:10	**0.2837 (0.003)**	0.2730 (0.011)	0.002
Area under AUC-abstention curve			
	BA	Proposed	p-Value
1:2	0.2545 (0.003)	**0.2673 (0.002)**	0.002
1:5	0.2350 (0.009)	**0.2587 (0.004)**	0.002
1:10	0.2491 (0.007)	**0.2770 (0.011)**	0.002
Area under G-mean-abstention curve			
	BA	Proposed	p-Value
1:2	0.2444 (0.005)	**0.2625 (0.005)**	0.002
1:5	0.1887 (0.017)	**0.2500 (0.008)**	0.002
1:10	0.2094 (0.011)	**0.2821 (0.010)**	0.002

and misdiagnosis. We presented a BUS diagnosis methodology which included image denoising, feature extraction, feature selection, and abstaining classification. The proposed bounded-abstaining classification method utilized AUC as the optimization target to accommodate imbalanced datasets. According to set conditions, the method rejected to classify some uncertain instances to ensure reliability.

The proposed method was validated using BUS datasets of different imbalance levels. The results showed that the AUC was less affected by the class imbalance than accuracy. Compared with an abstaining classification method, the proposed method yielded a better trade-off between performance and abstention with abstention at [0,0.3]; for fixed reject rates, the proposed method had significantly larger AUC and G-mean values. This indicates the potential practical value of the proposed method.

In practice, if some easy tumors can be diagnosed by CAD methods, this will alleviate doctors' workload. Based on the assumption, we proposed our BUS diagnosis methodology. In BUS diagnosis, the rejected examples deserve more attention, and domain experts should analyze these tumor images carefully. Hence, the abstaining classification method could help screen BUS images that are difficult to discriminate, thus saving time and enhancing the efficiency of practitioners. In addition, the limitation of the study is the small sample size of the imbalanced datasets. If a large number of instances are obtained, we believe that the performance-abstention curves can be more stable, and the appropriate parameters can be chosen based on the curves.

Acknowledgment

We appreciate the help of the radiologists from the First Affiliated Hospital of Harbin Medical University, China.

References

Abdel-Nasser, M., Melendez, J., Moreno, A., Omer, O.A. and Puig, D. (2017). Breast tumor classification in ultrasound images using texture analysis and super-resolution methods, *Engineering Applications of Artificial Intelligence* **59**: 84–92.

Acharya, U.R., Ng, W.L., Rahmat, K., Sudarshan, V.K., Koh, J.E., Tan, J.H., Hagiwara, Y., Yeong, C.H. and Ng, K.H. (2017). Data mining framework for breast lesion classification in shear wave ultrasound: A hybrid feature paradigm, *Biomedical Signal Processing and Control* **33**: 400–410.

Cai, L., Wang, X., Wang, Y., Guo, Y., Yu, J. and Wang, Y. (2015). Robust phase-based texture descriptor for classification of breast ultrasound images, *Biomedical Engineering Online* **14**(1): 26.

Chang, J.M., Moon, W.K., Cho, N., Yi, A., Koo, H.R., Han, W., Noh, D.-Y., Moon, H.-G. and Kim, S.J. (2011). Clinical application of shear wave elastography (SWE) in the diagnosis of benign and malignant breast diseases, *Breast Cancer Research and Treatment* **129**(1): 89–97.

Chen, S.-C., Cheung, Y.-C., Su, C.-H., Chen, M.-F., Hwang, T.-L. and Hsueh, S. (2004). Analysis of sonographic features for the differentiation of benign and malignant breast tumors of different sizes, *Ultrasound in Obstetrics and Gynecology* **23**(2): 188–193.

Cheng, H.-D., Shan, J., Ju, W., Guo, Y. and Zhang, L. (2010). Automated breast cancer detection and classification using ultrasound images: A survey, *Pattern Recognition* **43**(1): 299–317.

Daoud, M.I., Bdair, T.M., Al-Najar, M. and Alazrai, R. (2016). A fusion-based approach for breast ultrasound image classification using multiple-ROI texture and morphological analyses, *Computational and Mathematical Methods in Medicine* **2016**: 6740956.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers, *Machine Learning* **31**(1): 1–38.

Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letters* **27**(8): 861–874.

Fischer, L., Hammer, B. and Wersing, H. (2015). Efficient rejection strategies for prototype-based classification, *Neurocomputing* **169**: 334–342.

Fu, J., Li, Y., Li, N. and Li, Z. (2018). Comprehensive analysis of clinical utility of three-dimensional ultrasound for benign and malignant breast masses, *Cancer Management and Research* **10**: 3295–3303.

Gai, S., Zhang, B., Yang, C. and Yu, L. (2018). Speckle noise reduction in medical ultrasound image using monogenic wavelet and Laplace mixture distribution, *Digital Signal Processing* **72**: 192–207.

Garcia-Closas, M. *et al.* (2008). Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics, *PLoS Genetics* **4**(4): e1000054.

Guan, H., Zhang, Y., Cheng, H., Xian, M. and Tang, X. (2019). Ba2cs: Bounded abstaining with two constraints of reject rates in binary classification, *Neurocomputing* **357**: 125–134.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *Journal of Machine Learning Research* **3**(Mar): 1157–1182.

Haralick, R.M., Shanmugam, K. and Dinstein, I. (1973). Textural features for image classification, *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**(6): 610–621.

Hofmann, T. (1999). Probabilistic latent semantic analysis, *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden*, pp. 289–296.

Hong, X., Chen, S. and Harris, C.J. (2007). A kernel-based two-class classifier for imbalanced data sets, *IEEE Transactions on Neural Networks* **18**(1): 28–41.

Kang, S., Cho, S., Rhee, S.-j. and Yu, K.-S. (2017). Reliable prediction of anti-diabetic drug failure using a reject option, *Pattern Analysis and Applications* **20**(3): 883–891.

Lee, J., Nishikawa, R.M., Reiser, I. and Boone, J.M. (2018). Relationship between computer segmentation performance and computer classification performance in breast CT: A simulation study using RGI segmentation and LDA classification, *Medical Physics* **45**(8): 3650–3656.

Li, L., Zhou, X., Zhao, X., Hao, S., Yao, J., Zhong, W. and Zhi, H. (2017). B-mode ultrasound combined with color Doppler and strain elastography in the diagnosis of non-mass breast lesions: A prospective study, *Ultrasound in medicine & biology* **43**(11): 2582–2590.

Liberman, L. and Menell, J.H. (2002). Breast imaging reporting and data system (BI-RADS), *Radiologic Clinics of North America* **40**(3): 409–430.

Lin, C.-M., Hou, Y.-L., Chen, T.-Y. and Chen, K.-H. (2014). Breast nodules computer-aided diagnostic system design using fuzzy cerebellar model neural networks, *IEEE Transactions on Fuzzy Systems* **22**(3): 693–699.

Liu, Y., Cheng, H., Huang, J., Zhang, Y., Tang, X., Tian, J.-W. and Wang, Y. (2012). Computer aided diagnosis system for breast cancer based on color Doppler flow imaging, *Journal of Medical Systems* **36**(6): 3975–3982.

López, V., Fernández, A., García, S., Palade, V. and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, *Information Sciences* **250**: 113–141.

Monticciolo, D.L., Newell, M.S., Hendrick, R.E., Helvie, M.A., Moy, L., Monsees, B., Kopans, D.B., Eby, P.R. and Sickles, E.A. (2017). Breast cancer screening for average-risk women: Recommendations from the ACR commission on breast imaging, *Journal of the American College of Radiology* **14**(9): 1137–1143.

Moon, W.K., Chen, I.-L., Yi, A., Bae, M.S., Shin, S.U. and Chang, R.-F. (2018). Computer-aided prediction model for axillary lymph node metastasis in breast cancer using tumor morphological and textural features on ultrasound, *Computer Methods and Programs in Biomedicine* **162**: 129–137.

Mousania, Y. and Karimi, S. (2019). Contrast improvement of ultrasound images of focal liver lesions using a new histogram equalization, *in* K.S. Montaser (Ed.), *Fundamental Research in Electrical Engineering*, Springer, Singapore, pp. 43–53.

Pietraszek, T. (2007). On the use of ROC analysis for the optimization of abstaining classifiers, *Machine Learning* **68**(2): 137–169.

Prati, R.C., Batista, G. and Monard, M.C. (2011). A survey on graphical methods for classification predictive performance evaluation, *IEEE Transactions on Knowledge and Data Engineering* **23**(11): 1601–1618.

Rahmawaty, M., Nugroho, H. A., Triyani, Y., Ardiyanto, I. and Soesanti, I. (2016). Classification of breast ultrasound images based on texture analysis, *International Conference on Biomedical Engineering (IBIOMED), Yogyakarta, Indonesia*, pp. 1–6.

Rawashdeh, M., Lewis, S., Zaitoun, M. and Brennan, P. (2018). Breast lesion shape and margin evaluation: BI-RADS based metrics understate radiologists' actual levels of agreement, *Computers in Biology and Medicine* **96**: 294–298.

Rodriguez-Cristerna, A., Gomez-Flores, W. and de Albuquerque Pereira, W.C. (2018). A computer-aided diagnosis system for breast ultrasound based on weighted bi-rads classes, *Computer Methods and Programs in Biomedicine* **153**: 33–40.

Roffo, G., Melzi, S., Castellani, U. and Vinciarelli, A. (2017). Infinite latent feature selection: A probabilistic latent graph-based ranking approach, *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy,* pp. 1407–1415.

Roffo, G., Melzi, S. and Cristani, M. (2015). Infinite feature selection, *Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile*, pp. 4202–4210.

Roy, R., Ghosh, S. and Ghosh, A. (2018). Speckle de-noising of clinical ultrasound images based on fuzzy spel conformity in its adjacency, *Applied Soft Computing* **73**: 394–417.

Shan, J., Alam, S.K., Garra, B., Zhang, Y. and Ahmed, T. (2016). Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods, *Ultrasound in Medicine and Biology* **42**(4): 980–988.

Shi, X., Cheng, H.-D., Hu, L., Ju, W. and Tian, J. (2010). Detection and classification of masses in breast ultrasound images, *Digital Signal Processing* **20**(3): 824–836.

Simeone, P., Marrocco, C. and Tortorella, F. (2012). Design of reject rules for ECOC classification systems, *Pattern Recognition* **45**(2): 863–875.

Singh, B.K., Verma, K., Panigrahi, L. and Thoke, A. (2017a). Integrating radiologist feedback with computer aided diagnostic systems for breast cancer risk prediction in ultrasonic images: An experimental investigation in machine learning paradigm, *Expert Systems with Applications* **90**: 209–223.

Singh, K., Ranade, S.K. and Singh, C. (2017b). A hybrid algorithm for speckle noise reduction of ultrasound images, *Computer Methods and Programs in Biomedicine* **148**: 55–69.

Tesfahun, A. and Bhaskari, D.L. (2013). Intrusion detection using random forests classifier with smote and feature reduction, *International Conference on Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), Pune, India*, pp. 127–132.

Tortorella, F. (2000). An optimal reject rule for binary classifiers, *in* F.J. Ferri *et al.* (Eds), *Advances in Pattern Recognition, SSPR/SPR 2000*, Lecture Notes in Computer Science, Vol. 1876, Springer, Berlin/Heidelberg, pp. 611–620.

Tortorella, F. (2004). Reducing the classification cost of support vector classifiers through an ROC-based reject rule, *Pattern Analysis and Applications* **7**(2): 128–143.

Wang, Z., Wang, Z., He, S., Gu, X. and Yan, Z.F. (2017). Fault detection and diagnosis of chillers using Bayesian network merged distance rejection and multi-source non-sensor information, *Applied Energy* **188**: 200–214.

Wu, G. and Chang, E.Y. (2005). KBA: Kernel boundary alignment considering imbalanced data distribution, *IEEE Transactions on Knowledge and Data Engineering* **17**(6): 786–795.

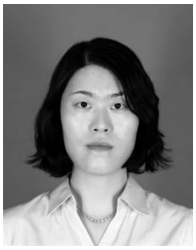
Yassin, N.I., Omran, S., El Houby, E.M. and Allam, H. (2017). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review, *Computer Methods and Programs in Biomedicine* **156**: 25-45.

Yu, H. and Ni, J. (2014). An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **11**(4): 657–666.

Yu, X., Hao, X., Wan, J., Wang, Y., Yu, L. and Liu, B. (2018). Correlation between ultrasound appearance of small breast cancer and axillary lymph node metastasis, *Ultrasound in Medicine & Biology* **44**(2): 342–349.

Zhang, J., Lin, G., Wu, L., Wang, C. and Cheng, Y. (2015). Wavelet and fast bilateral filter based de-speckling method for medical ultrasound images, *Biomedical Signal Processing and Control* **18**: 1–10.

Zhou, S., Shi, J., Zhu, J., Cai, Y. and Wang, R. (2013). Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image, *Biomedical Signal Processing and Control* **8**(6): 688–696.



Hongjiao Guan received her PhD degree from the School of Computer Science and Technology, Institute of Technology, Harbin, China, in 2020. Now she is a lecturer at the School of Cyber Security, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. Her research interests mainly include pattern recognition and machine learning.

Yingtao Zhang received her MS degree from the Computer Science School of the Harbin Institute of Technology, China, in 2004, and her PhD degree in pattern recognition and intelligence system from the Harbin Institute of Technology in 2010. Now, she is an associate professor at the School of Computer Science and Technology, Harbin Institute of Technology. Her research interests include pattern recognition, computer vision and medical image processing.

Heng-Da Cheng received his PhD degree in electrical engineering from Purdue University West Lafayette in 1985. Now, he is a full professor at the Department of Computer Science and an adjunct full professor at the Department of Electrical Engineering, Utah State University, Logan, USA. His research interests include image processing, pattern recognition, computer vision, artificial intelligence, medical information processing, fuzzy logic, genetic algorithms, neural networks, parallel processing, parallel algorithms, and VLSI architectures. Doctor Cheng is also an associate editor of *Pattern Recognition*, *Information Sciences* and *New Mathematics and Natural Computation*.

Xianglong Tang is a professor and a PhD supervisor at the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include artificial intelligence and information processing.

Received: 30 March 2019
Revised: 28 October 2019
Re-revised: 10 December 2019
Accepted: 20 December 2019