Computers in Biology and Medicine 83 (2017) 157–165

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning

Wei Lu, Zhe Li, Jinghui Chu*

School of Electronic Information Engineering, Tianjin University, Tianjin 300072, PR China

ARTICLE INFO

Keywords:
Computer-aided diagnosis
Breast cancer
MRI
Feature selection
Classification
Ensemble learning

ABSTRACT

Breast cancer is a common cancer among women. With the development of modern medical science and information technology, medical imaging techniques have an increasingly important role in the early detection and diagnosis of breast cancer. In this paper, we propose an automated computer-aided diagnosis (CADx) framework for magnetic resonance imaging (MRI). The scheme consists of an ensemble of several machine learning-based techniques, including ensemble under-sampling (EUS) for imbalanced data processing, the Relief algorithm for feature selection, the subspace method for providing data diversity, and Adaboost for improving the performance of base classifiers. We extracted morphological, various texture, and Gabor features. To clarify the feature subsets' physical meaning, subspaces are built by combining morphological features with each kind of texture or Gabor feature. We tested our proposal using a manually segmented Region of Interest (ROI) data set, which contains 438 images of malignant tumors and 1898 images of normal tissues or benign tumors. Our proposal achieves an area under the ROC curve (AUC) value of 0.9617, which outperforms most other state-of-the-art breast MRI CADx systems. Compared with other methods, our proposal significantly reduces the false-positive classification rate.

1. Introduction

Breast cancer is a disease that is caused by malignant cells in the breast tissues [1]. According to American National Cancer Institute, the incidence rate of female breast cancer was 124.8 per 100,000 women per year from 2008 to 2012. In 2015, it is estimated that there were 231,840 new breast cancer cases and about 40,290 people died from this disease. In the U.S., female breast cancer is responsible for 14% of all new cancer cases [2], indicating that female breast cancer is the most common cancer. The mortality rate from breast cancer is also the highest among women [3]. Breast cancer pathogenesis remains unknown, and there is no effective way to prevent this disease [4]. However, early detection and diagnosis of breast cancer can help to significantly reduce the mortality rate. Thus, much research has been performed on the early detection of malignant masses. Modern imaging techniques including ultrasound, mammography, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) have been widely used for the early detection and diagnosis of breast cancer [5]. Among these techniques, MRI is well-known for its superiority in prognosis, diagnostic accuracy, staging, and preoperative planning [67]. It was also shown that MRI has better sensitivity than mammography and MRI diagnostic results are only minimally influenced by breast density [8]. Therefore, MRI is considered to be an important tool in breast cancer clinical diagnosis [5].

The goal of Computer-Aided Detection and Diagnosis (CAD) is to achieve a high diagnostic sensitivity for breast cancer and to maintain a low the false positive classification (FPC) rate [5]. In this process, Computer-Aided Diagnosis (CADx) is regarded as a key technique to reduce the FPC rate [9].

An important task of CADx is to make an accurate mass classification and decide whether a region of interest (ROI) is malignant. Enough high-quality features that characterize malignant masses are needed for training the classification model and for class prediction [10]. Thus, many features, such as morphological features, texture features, and frequencial features, have been extracted and used widely in many studies [11,12,13,14]. However, some extracted features may be redundant and irrelevant to the classification task. In addition, too many input features may increase the computational complexity and cause the curse of dimensionality, thereby significantly diminishing the diagnostic accuracy. Thus, feature selection plays a crucial role in improving CADx system performance. Genomic algorithm (GA) [15] and support vector machine-based recursive feature elimination (SVM-RFE) [9] have been adopted in the feature selection process for CADx systems.

* Corresponding author.
E-mail addresses: luwei@tju.edu.cn (W. Lu), tywzlizhe29121@126.com (Z. Li), cjh@tju.edu.cn (J. Chu).

http://dx.doi.org/10.1016/j.compbiomed.2017.03.002
Received 30 September 2016; Received in revised form 25 February 2017; Accepted 1 March 2017

In addition to acquiring representative features of breast masses, it is also important to build and train a robust classifier. According to our knowledge, most popular classifiers are designed under the assumption that the data set used for training is balanced, which means that the number of samples in the majority class is similar to that in the minority. However, this prerequisite is difficult to achieve in CADx. There are usually much fewer images with malignant masses than those without masses or with only benign masses [16]. This may reduce the diagnostic sensitivity, and malignant masses are likely to be wrongly classified as being normal. Consequently, it is necessary to alleviate the influence caused by data imbalance. Ensemble learning algorithms such as Ensemble of Under-sampled SVM (EUS-SVM) [17] and RUSBoost [18] have been shown to perform well in breast cancer CAD [16] and other medical classification systems [19].

In this paper, we propose a MRI CADx system focusing on diagnosis of malignant breast masses, based on feature selection and ensemble learning. First, morphological, texture, and Gabor features were extracted to characterize breast cancer masses. We then used the Relief algorithm [20] to find the optimal feature subset for the classifier training. These features were then fed to a novel ensemble learning framework based on the combination of EUS and the subspace technique. The experimental results indicate that our proposal outperforms the other state-of-the-art methods in diagnostic sensitivity, but the FPC rate increases slightly.

The main contributions of this paper can be summarized as follows:

1. The dimensionality of the features we use is larger than most state-of-the-art methods. Various features including morphological, Gabor, and several types of texture features were extracted to comprehensively characterize breast masses.
2. We selected the optimal feature subset from the original feature set using Relief, based on their type, which helps reduce the redundant and irrelevant features and takes the physical meaning of features into consideration.
3. We propose a novel ensemble learning framework based on the combination of EUS, subspace, and Adaboost, which helps to alleviate the data imbalance problem and improves the overall classification accuracy of the CADx system.

The remainder of this paper is organized as follows: Section 2 describes our methodology in detail; Section 3 introduces our data set, the experimental setup, and our performance evaluation metrics; Section 4 presents the experimental results and demonstrates the effectiveness of each individual component of our proposal; Section 5 discusses the reason for our system's superiority; and Section 6 includes concluding remarks (Fig. 1).

2. Methodology

In this section, we discuss the methodology used in our proposal. The structure of our CADx system is shown in Fig. 2. Because automated ROI segmentation may cause some errors [21], here, we segmented the ROIs from breast MRI images manually with the aid of physicians' marks. In subsection A, we present the features that we extracted for characterization of breast cancer images. In subsection B, we introduce our method for feature selection and the process of subspace construction. In subsection C, we briefly introduce our ensemble learning framework for imbalanced data processing. In subsection D, we introduce our base classification algorithm. Finally, in subsection E, we present the pseudo-code for our proposal.

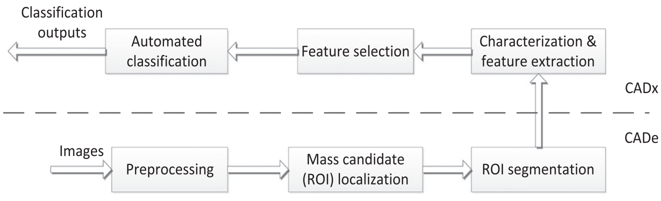

Fig. 1. This is a diagram of CAD scheme. The lower part below the dashed line shows the components of CADe, and the upper part above the dashed line shows the components of CADx.

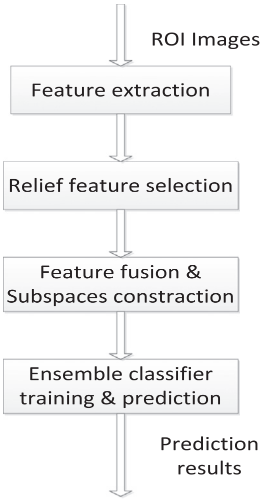
ROI Images
Feature extraction
Relief feature selection
Feature fusion & Subspaces constraction
Ensemble classifier training & prediction
Prediction results

Fig. 2. The general structure of our CADx system.

2.1. Feature extraction

In the pattern recognition and image processing fields, the aim of feature extraction is to acquire numeric variables that can reflect the intrinsic image properties [22]. Malignant breast masses usually have a spiculated, rough, and blurry boundary, while benign masses usually have a round, smooth, and well-defined boundary [23]. These morphological features are crucial to the classification. However, not all morphology information is suitable for describing the ROI images. Texture and frequencial features that focus more on the tissue composition are also useful and effective for characterizing breast masses. In this study, we extract various types of features, including morphological features, gray level co-occurrence matrix (GLCM, also known as Haralick features [24]), gray level difference matrix (GLDM), gray level run length matrix (GLRLM), gradient-gray level co-occurrence matrix (GGLCM) and Gabor features, for the additional classification task. Among these features, GLCM, GLDM, GLRLM, and GGLCM features belong to the texture statistical feature family, and Gabor features belong to the frequency domain feature family. A brief introduction to these features is provided below.

2.1.1. Morphological features

Morphological features are variables that describe the shape, edge, and geometric properties of masses [25]. In our proposal, circularity, mean and standard deviation of the radial length, eccentricity, entropy of the intensity distribution, mean and standard deviation of the intensity, area, mean and standard deviation of the fractal dimension index, inertial momentum, anisotropy, entropy of the contour gradient, smoothness, skewness and kurtosis were extracted as morphological features. The dimensionality of our implementation in this part is 16.

2.1.2. Haralick features

Haralick features [24] are statistical texture features based on GLCM. Unlike the gray level histogram which can only reflect the gray level and its spatial distribution, GLCM can display information about the relative position of pixel pairs with different distances and angular relationships in the image [26]. For each breast MRI image, we extracted the following 11 features from its GLCM; namely angular second moment, contrast, correlation, difference moment, homogeneity (inverse difference moment), sum average, sum variance, sum entropy, entropy, difference variance, and difference entropy. For a more detailed characterization of the image, we extracted Haralick features in 16 directions and distances rather than just from the traditional four directions with fixed distances. Therefore, our original Haralick feature set for classification contains 176 features.

2.1.3. GLDM features

In grayscale images, the difference in the gray level between neighboring pixels can reflect the rate of energy change of the electromagnetic wave radiation [27]. As a result, the gray level difference can provide a description of the image texture. The gray level difference can be defined as (1):

$$\Delta f = f(x + \Delta x, y + \Delta y) - f(x, y), \tag{1}$$

where $(\Delta x, \Delta y)$ is the displacement vector of the pixel pairs. For a specific displacement vector, all possible gray level difference values can be denoted by the probability density function, and a gray level difference histogram can be acquired. Four features (contrast, angular second moment, mean, and entropy) can be extracted from the histogram. In our proposal, we used 16 different displacement vectors and thus extracted 64 GLDM features in total.

2.1.4. GLRLM features

In the image processing field, gray level run length is defined as the number of consecutive pixels that have the same grayscale level in one direction [28]. The texture of one image can be characterized by the gray level, length, and direction of the run in GLRLM features. Because there was no difference between 45° and 135° in GLRLM, we calculated three matrices at 0°, 45°, and 90°. Short run emphasis, long run emphasis, gray level distribution, run length distribution, and run percentages were extracted from all matrices. Thus, we obtained a total of 15 GLRLM feature.

2.1.5. GGLCM features

GGLCM makes use of both gradient and gray level information to characterize the texture of images [29]. In this way, both the edge and the interior part of one image can be taken into consideration. To make the features, we extracted effective and precise descriptions of the ROIs and reduced the computational cost; we did not use the original 256-level gradient and grayscale measurements. Instead, we detected the edges using the Roberts operator and then set the number of gradient levels to 8, to obtain the gradient image G. We also set the number of grayscale levels to 16, to obtain the grayscale image F. The gradient-gray level co-occurrence matrix can be defined as (2):

$$H(i, j) = count(F(m, n) = i, G(m, n) = j), \tag{2}$$

where (m,n) is the coordinate of the current pixel in both F and G. In our setting, the gradient-gray level co-occurrence matrix H was an 8×16 matrix. For one image, we extracted the small gradient dominance, big gradient dominance, grayscale asymmetry, gradient asymmetry, energy, grayscale mean, gradient mean, grayscale variance, gradient variance, correlation, grayscale entropy, gradient entropy, grayscale-gradient mixed entropy, inertia, and difference moment from the GGLCM as features.

2.1.6. Gabor features

The Gabor wavelet is widely applied in visual comprehension. It is sensitive to the edge of images, which can provide good direction and scale selection characteristics. It is not sensitive to illumination changes. Gabor features are based on a group of wavelets at different directions and frequencies [13]. To lower the computational cost while retaining the accuracy of the features, we set four frequencies and eight directions, thus generating 32 different Gabor wavelet filters. We calculated the convolution of the original image and each Gabor filter to obtain a feature image, and then extracted the mean and standard deviation of the pixels in the feature image as features. We used 64 Gabor features in total.

2.2. Feature selection

Feature selection plays an important role in training classifiers. Assuming that the original feature space contains D features, the goal of the feature selection process is to find an optimal subset that contains only d ($d \le D$) features. Feature selection can reduce the computational complexity and improve the classification accuracy. Feature selection methods can be mainly categorized into two types: wrapper and filter [30]. Wrapper is in combination with specific classifiers. The performance and importance of features are evaluated using a specific classification algorithm. Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), and Sequential Floating Selection (SFS) [31] are representative wrapper methods. Filter does not consider specific classification algorithms and gets each feature's contribution directly from the original data. Generally, filter methods are faster than wrapper methods and they are not sensitive to the classifier. Among filter methods, Relief [20], which is based on feature weighting, is widely regarded as an effective filter selection method. Relief is easy to implement and has lower computational complexity, and therefore we used it for feature selection. The Relief algorithm can be described as follows:

1. The training was set as D, and the weight of each feature was initialized as 0. For each instance R in D, its nearest neighbor instance with the same class as NH (NearHit) and the nearest instance in the different class from NM (NearMiss) was found based on the Euclidean distance measure.
2. An instance R was randomly sampled with replacement in D for m times. For each feature i, the weight of w_i was updated as follows:

$$w_i = w_i - dif(i, R, NH)/m + dif(i, R, NM)/m, \tag{3}$$

where dif is the difference between D and its NearHit or NearMiss instance, and can be computed as follows:

$$dif(i, R, N) = \left|\frac{R - N}{\max(i) - \min(i)}\right|. \tag{4}$$

Thus, features that can distinguish samples and their NearMiss neighbors are regarded as more relevant and can be kept in the final optimal feature subset.

3. The final weight of each feature can be acquired after m rounds of iteration, then these weights are compared with a preset value, th. Features that have a weight higher than th will be added to the final feature subset.

Among all of the features extracted for training and testing, morphological features, GLRLM features, and GGLCM features are excluded from the feature selection process. Because morphological features always play an important role in clinical diagnosis, they are usually of high value and more attention is paid to them. Because GLRLM and GGLCM features both contain a limited number of features, it is not necessary to make dimensionality reductions on these features. To preserve the physical meaning of features, features are not mixed together before feature selection. Instead, the remaining types of features are selected using Relief, and therefore Haralick, GLDM, and Gabor feature subsets are formed.

In our proposal, the parameters mentioned in this subsection are fixed. As seen from Section 2.1, we extracted 339 features in total, so D is 339. Regarding other parameters in the Relief algorithm, we set m to 1000 and the threshold th to 800, which means that the average threshold is 0.8 for each feature dimension to be selected for the training of classifiers.

2.3. Imbalanced data processing

Data imbalance is a common problem in medical classification tasks. In breast CADx systems, ROIs are segmented automatically or manually; only a small part of the ROIs contain malignant tumors and most ROIs are false-positive ROI (FPROI) instances. Considering that most binary classification algorithms were designed under the assumption that there is no significant difference between the size of the minority (positive) and that of the majority (negative) class, the performance of classifiers might be seriously decreased as a result of data imbalance. The positive area is likely to be eroded by the negative area and thus the true positive rate will decrease. To solve this problem, we propose an ensemble method that combines undersampling, subspace, and boosting based on the guideline of bias-variance decomposition.

2.3.1. Bias-variance decomposition

For a trained classification model, bias-variance decomposition [32] can be used for analyzing its prediction error and generalization ability. The generalization error can be broken down into three components: bias, variance, and noise as follows:

$$error = bias^2 + variance + noise. \tag{5}$$

Bias measures the difference between the desired output and the real label. Variance is used to measure the inner deviation of the testing set. It can be characterized as the deviation between the outputs and their average. Noise is the lower boundary of the generalization error and measures the difficulty of the classification task. Ensemble methods can make combinations of multiple classifiers and thus reduce the generalization error in different ways. Bagging-based methods aim at reducing the variance of base classifiers by undersampling, while Boosting-based methods focus on reducing the bias by weighting strategies. Additionally, for base classifiers, they should be accurate and have good diversity. Subspace strategy divides the feature vector space into several spaces and trains classifiers with these subsets. Thus, the diversity of classifiers is ensured and their ensemble can obtain better performance. Here, we combine the feature subspace, Bagging, and Boosting to build an ensemble framework to address the data imbalance problem.

2.3.2. Variance reduction—ensemble of undersampling

EUS [17] is a bagging-based strategy that can help to reduce the generalization variance. Unlike traditional bagging algorithms, EUS samples the negative instances without replacement to build negative (i.e. false-positive classification) subsets, all of which are a similar size as that of the positive class. Each negative subset was combined with all positive instances to build several balanced subsets for training. In the testing process, one testing instance was classified by all sub-classifiers and the outputs were combined using a pre-determined rule, such as majority voting. Compared with random undersampling, which only uses a small part of negative instances, EUS can make use of all negative instances to avoid losing representativeness. The general structure of EUS is shown in Fig. 3.

2.3.3. Diversity enhancement—feature subspace method

The training instances are usually characterized by a set of features. Different feature subspaces (i.e. feature subsets) can provide different perspectives to observe the data. As a result, it will form different classification models after training, which can increase the feature disturbance and thus help improve the base classifier diversity. In our proposal, for each balanced subset acquired from EUS, we use the feature subspace method to train different classifiers. Morphological features will be fused and directly put together with five other types of features, as mentioned in Section 2.1, to build five different feature subsets. Among these features, Haralick, GLDM, and Gabor features are the optimal feature subsets selected by Relief. Thus, for each balanced EUS subset, there will be five different base classifiers. This process is described in Algorithm 1 as the *feature_combination* function.

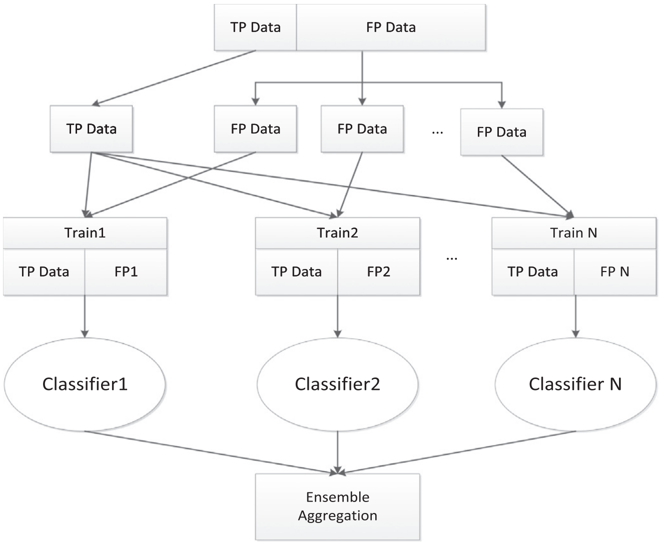

Fig. 3. The general structure of EUS. *N* represents for the imbalanced ratio. It is approximately equal to the ratio of the size of negative class to that of the positive class. The ensemble aggregation strategy can be either majority voting, weighted voting, function value aggregation or other effective aggregation methods.

2.3.4. Bias reduction—adaboost

Boosting is an iterative procedure for changing the distribution of training instances adaptively to help base classifiers focus more on the instances, which are more likely to be misclassified [33]. The weights of the instances that are hard to classify will increase during the iteration until they can be correctly classified in subsequent iterations. Adaboost [34] is one of the most well-known implementations of Boosting algorithms. It aims to reduce the output of the logarithmic loss function. The weight of a base classifier is related to its error rate and the weight of each training instance changes according to its classification accuracy. The generalization bias can drop continuously in the iteration process. In our proposal, Adaboost is combined with the feature subspace method for improving the performance of each base classification model trained by one feature subspace.

2.3.5. Result aggregation—majority voting

For ensemble learning methods, different result aggregation methods may cause different classification outputs. Majority voting, weighted voting, and function value aggregation are three widely used methods to make the aggregation. However, experimental results have shown that there is no significant difference in the classification performance among these methods [17]. In addition, weighted strategy and functional value aggregation can cause overfitting [35]. Thus, we chose majority voting to aggregate the subspace classifiers' outputs and EUS outputs.

2.4. Classification method

In our proposal, we chose C4.5 [36] as the base classifier for our classification task. C4.5 is a decision tree that chooses appropriate features and nodes based on the information gain ratio. It is a weak learner and can show great performance in ensemble methods. C4.5 is a white box, which means it can generate clear classification rules. Compared with ID3 [37], another classic decision tree algorithm that is based only on information gain, C4.5 tends to choose the easier and shorter classification paths. Thus, C4.5 is in accordance with the principle of Occam's razor. C4.5 is among the top 10 data mining algorithms [38] and has been widely used in imbalanced classification tasks [39,19].

2.5. Overall procedure

In summary, our proposal for addressing the data imbalance problem is an ensemble method that combines EUS, the feature subspace method, and Adaboost to reduce the generalization error. We assume that there are n instances for training in total. Each sample is represented as (x_i, y_i). Here, x_i is the feature vector of one sample, and y_i, which can be either +1 (positive) or −1 (negative), is its class label, and the number of positive instances in the training set is m. Haralick, GLDM, GLRLM, GGLCM, and Gabor features are labeled from 1 to 5. The entire imbalanced data processing procedure can be described as Algorithm 1.

3. Experimental setup

In this section, we introduce our experimental framework and parameter settings. In subsection A, we give a brief description to our breast MRI data set. In subsection B, we introduce the parameters and other settings of C4.5. In subsection C, we present the metrics we use to evaluate the performance of our proposal.

3.1. Data set

Although many breast MRI CADx systems define the ROIs automatically, we selected ROIs manually based on physicians' annotations to avoid potential segmentation errors, as described by Abdel-Nasser et al. [21]. Fig. 4 shows some manually segmented ROIs from the data set.

Although our data set was built manually, we made it imbalanced to mimic the real condition. The data set contains 438 positive instances and 1898 negative instances in total, and the imbalanced ratio is approximately 4. The MR images were collected from different breast cancer MRI volumes at Shandong Cancer Hospital, a top cancer diagnosis and therapy hospital in China. Images were generated on a Philips 3.0 TMR system at a resolution of 360 bpi. The patients were from different regions throughout mainland China, and the patients' age ranged from 35 to 60, which means that the data set is a representative subset of all breast cancer patients. For each patient, multiple images were collected in our data set. However, the 2D MRI slices were taken from different positions and angles, so these images are different even though they are taken from a limited number of patients (i.e. 54 patients in total). Because of these measures, the representativeness of our data set can be ensured. In the EUS framework, we built four balanced subsets for subspace training and testing, and we set the majority voting threshold to 3. In each subset, we set the voting threshold for the five subspace classifiers to 3.

The negative set mainly consists of mass-like lesions and edemas shown in the original images, but the number of benign masses is not large (less than 50); and the data set does not contain any ductal carcinoma in situ (DCIS). This is because MR technology is typically used after the detection of positive masses, and the images we acquired are all from positive masses.

3.2. Classification algorithm settings

In our proposal, C4.5 was chosen as the base classifier. We used the non-pruning tree and set the confidence level as 0.25. The number of patterns per leaf should be no less than 2. In the Adaboost part, we set the number of iterations to 25. All parts of our proposal were implemented and executed on Matlab 2014a.

3.3. Performance metrics

We measured our proposal using various representative performance metrics. We applied sensitivity, specificity, and the FPC rate to evaluate the performance from one specific perspective. Additionally, the $F_1 - measure$ and geometric mean (G-mean) were used for characterizing the trade-off between the TPC and FPC rate. The definitions of these metrics are described below:

$$sensitivity = \frac{TPC}{TPC + FNC}, \tag{6}$$

$$specificity = \frac{TNC}{FPC + TNC}, \tag{7}$$

$$FPC\ rate = \frac{FPC}{FPC + TNC}, \tag{8}$$

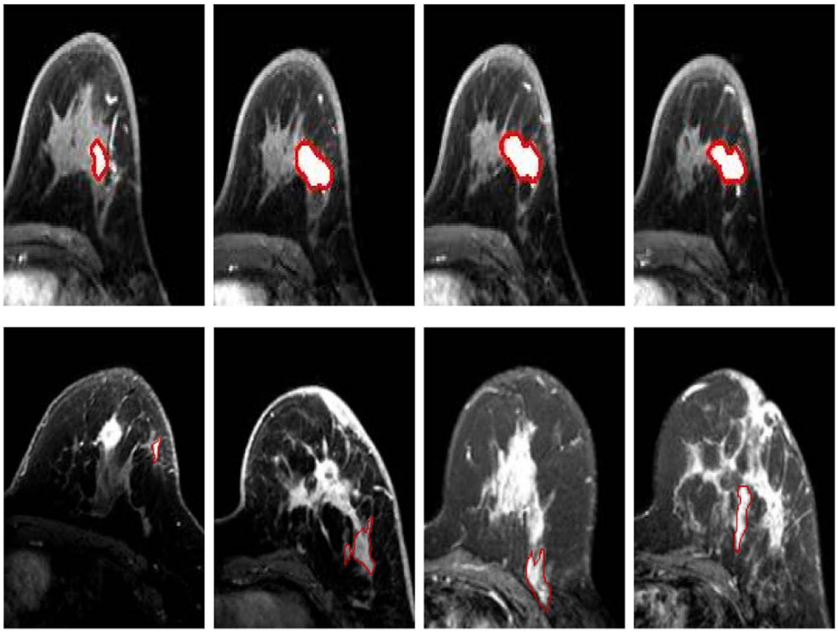

Fig. 4. Here are some examples in our data set. The ROI selected in each image is indicated with a red curve. The upper 4 images show malignant masses (i.e. positive instances), and the lower 4 images show false positive ROI areas (i.e. negative instances) with no masses or just benign masses.

$$F_1 - measure = \frac{2TPC}{2TPC + FPC + FNC}, \quad (9)$$

$$G - mean = \sqrt{sensitivity \times specificity}. \quad (10)$$

Here, *TNC* and *FNC* represent the number of correctly diagnosed negative cases and the wrongly diagnosed positive cases, respectively.

In addition to the metrics above, the receiver operating characteristic curve (ROC), a graph-based technique to visualize the classifiers' organization and performance, was used [40]. ROC was introduced to machine learning analysis from the field of medical decision making and it can depict the trade-off between the TPC rate and FPC rate more efficiently and intuitively. Additionally, area under the ROC curve (AUC) is based on the ROC, which provides a numerical measurement on the classifier's performance. A good classifier often achieves a high AUC value, which means that the classifier has a high TPC rate and a low FPC rate at the same time. We used a ten-fold cross-validation strategy in our experiments. In this situation, the sampling process is performed separately in the positive and negative sets. Both sets are divided into 10 subsets, and each positive subset is combined with a negative subset. Thus, we had 10 imbalanced subsets for the ten-fold cross-validation.

4. Experimental results and empirical comparisons

In this section, we present the experimental results of our breast MRI CADx system and illustrate its superiority. In subsection A, we present the overall experimental results with several performance metrics mentioned in Section 3.3. In Section 4.2, we compare our ensemble framework with its individual components, including EUS, Relief, and the feature subspace method, to show the effectiveness of our strategy. Finally, in Section 4.3, to demonstrate the superiority of our proposal, we also make some empirical comparisons with several state-of-the-art breast cancer CADx systems.

4.1. Overall results

4.1.1. Numerical results

Among the performance metrics mentioned in Section 3.3, sensitivity, specificity, $F_1 - measure$, G-Mean, and FPC rate describe the performance numerically. The numerical results of our proposal are shown in Table 1.

4.1.2. ROC analysis

The ROC curve of our proposal is shown in Fig. 5 and the value of AUC is 0.9617.

4.2. Ensemble strategy vs. individual components

As mentioned above, our proposal mainly consists of four components: feature selection (with Relief algorithm), subspace method, Adaboost, and imbalanced data processing (with EUS). To demonstrate the effectiveness of our ensemble strategy, we compared our ensemble method with four methods in which one component is removed. We generated four comparative methods by removing components from the proposed framework one at a time. Table 2 gives a brief description of the methods for comparison.

The experimental results are shown in Table 3. The ensemble strategy is better than nearly all the other methods, as assessed using these metrics.

Table 1
The experimental results of our proposal (/%).

Sensitivity	Specificity	$F_1 - measure$	G-mean	FPC rate
90.19	96.31	87.46	93.15	3.69

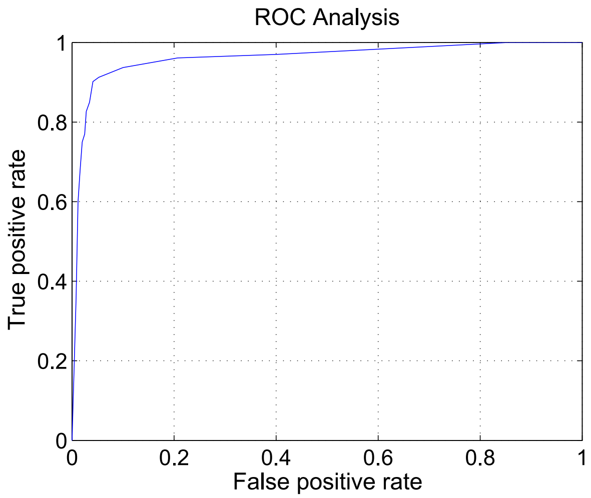

Fig. 5. The ROC graph of our proposal.

Table 2
Information on frameworks used for comparison.

	Relief	Subspace method	Adaboost	EUS
Ensemble strategy	√	√	√	√
Without selection		√	√	√
Without subspace	√		√	√
Without boosting	√	√		√
Without EUS	√	√	√	

4.3. Empirical comparisons with state-of-the-art breast cancer CADx systems

In addition to making comparisons among components within our own framework, we also compared our proposal with several state-of-the-art breast CADx systems to classify masses in breast MRI images. Table 4 shows the comparison results. It is suggested that our proposal can achieve a lower FPC rate with only a small decrease in sensitivity.

Additionally, we performed a statistical comparison between our proposal and the framework designed in [41], which is the best of the frameworks in Table 4. For imbalanced data sets, because there are much more negative samples than the positive ones and thus the precision metric will greatly influenced by data distribution, $F_1 - measure$ will not be accurate enough. Thus, we use other dualistic metrics (i.e. G-mean, AUC) and FPC, which is also important for the an CADx system, for comparison. The p-value of a pairwise *t*-test between the two frameworks is less than 0.01, which means our proposal is significantly superior to other state-of-the-art breast cancer CADx frameworks.

5. Discussion

Except for the conclusions drawn above, the experimental results presented in Fig. 5, Table 3 and 4 and the related experimental results reveal several facts, as described below.

First, as indicated in Table 3, our proposal outperforms other control groups generated for comparison, which means that all of our four individual components can help promote the diagnostic accuracy. EUS, which is used for balancing the data distribution, helps to build the classification decision boundary correctly and avoids the positive boundary degradation caused by the data imbalance issue. Relief, which is used for the feature selection, helps with the dimensionality reduction and thus reduces the computational complexity, as well as extracts optimal feature subsets. The subspace method provides data diversity and trains multiple base classifiers, each of which is suitable for one specific feature space with a clear physical meaning. Finally,

Table 3
The experimental result on ensemble strategy vs. individual components.

	Sensitivity (/%)	Specificity (/%)	$F_1 - measure$ (/%)	G-mean (/%)	FPC rate (/%)	AUC
Ensemble strategy	90.19	96.31	87.46	93.15	3.69	0.9617
Without selection	86.53	96.26	85.45	91.24	3.74	0.9435
Without subspace	88.80	95.89	85.99	92.25	4.11	0.9441
Without boosting	84.93	95.47	83.03	90.00	4.53	0.9312
Without EUS	57.31	97.73	68.29	74.63	2.27	0.7752

Table 4
Comparison on our proposal and state-of-the-art breast MRI CADx systems.

	Sensitivity (/%)	Specificity (/%)	$F_1 - measure(/\%)$	G-mean (/%)	FPC rate (/%)	AUC
K. Nie et al. (2008) [42]						0.860
D. M. Renz et al. (2012) [41]	96.5	75.5	91.7	85.4	14.0	0.935
E. Honda et al. (2015) [43]	87.1	82.1	89.3	84.6	10.4	0.829
S. Song et al. (2015) [44]	100.0	77.6		88.1		0.888
our proposal	90.19	96.31	87.46	93.15	3.69	0.9617

Table 5
Comparison on different result aggregation methods.

	Sensitivity (/%)	Specificity (/%)	$F_1 - measure$	G-mean (/%)	FPC rate	AUC
majority voting	90.19	96.31	87.46	93.15	3.69	0.9617
weighted voting	89.49	96.05	86.69	92.69	3.95	0.9572
function value aggregation	88.62	96.15	86.23	92.24	3.85	0.9558

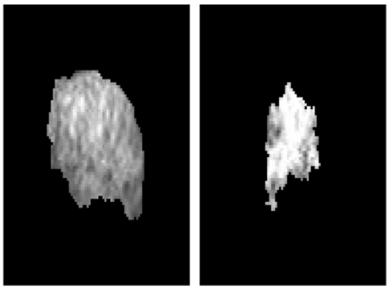

Fig. 6. Here are two examples misclassified images. Both of them are the edemas in the original images.

Adaboost helps to improve the classification ability of each C4.5 base classifier.

Second, data imbalance processing plays the most important role in our proposal. The overall performance has the most significant decrease without the EUS procedure. This is because EUS is the first step in our proposal and other methods can gain little effect if the data distribution is imbalanced [39]. Other individual parts are also helpful for improving the overall performance, but the benefit is not as large as that provided by EUS.

Third, for the part that aggregates the classification results, there is no significant difference among various popular aggregation methods, including weighted voting, function value aggregation, and the simplest majority voting [17], which we use in our proposal. Complicated aggregation methods may lead to overfitting and thereby reduce the generalizability of the results. Table 5 compares the performance of these three classifier aggregation methods. The statistical significance is also measured using a pairwise t-test. Among the three pairs available, the lowest p-value is 0.0431, which is much larger than the threshold that can be seen as significant (i.e. $p \leq 0.01$). Therefore, it is clear that there is no significant difference among these pairs, and the majority voting method is better than the other methods.

Fourth, after analyzing the experimental results, we found that we can achieve high accuracy on the recognition of malignant masses, and most of the misclassified samples are edema images, which are similar in morphology and texture to malignant masses. Two examples are shown in Fig. 6.

6. Concluding remarks

In this paper, we propose an automated breast MRI CADx system. Because our work focuses on improving the classification performance for malignant breast mass lesions using imbalanced data, we segmented breast ROIs manually rather than automatically, which excludes possible errors in automated segmentation. We propose an ensemble strategy that combines EUS, the feature subspace method, and boosting, as well as acquiring optimal feature subsets using Relief. The results and empirical comparisons have suggested that our ensemble proposal is better than its individual components. Additionally, it outperforms most state-of-the-art breast CADx systems.

Future work will continue to focus on improving the diagnostic accuracy of the breast MRI CADx system. We are preparing to train different kinds of base classifiers in the same system to enhance the diversity and see whether it helps to improve the overall diagnostic performance. In addition to the low-level image features we used in this paper, we will investigate higher-level features that are extracted by deep learning and other up-to-date techniques. We will investigate these features and try to obtain better feature vectors for classification. In addition, we will consider combining the breast MRI super-resolution technology with our proposal. Thus, we can obtain images of higher quality and improve the diagnostic accuracy.

Conflict of interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgements

Support for this research was provided by National Natural Science Foundation of China (NSFC) grant No. 61271069. We would like to thank Honghui Hou, Fengshuo Hu, and Zerui Wu for their contribution in building the data set used in this paper. We are also grateful to the editor and reviewers for their valuable comments and suggestions on this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2017.03.002.

References

[1] R. Lag, D. Harkins, M. Krapcho, A. Mariotto, B. Miller, E. Feuer, L. Clegg, M. Eisner, M. Horner, N. Howlader, et al., Seer cancer statistics review, 1975–2003, Bethesda, National Cancer Institute.

[2] N. Howlader, A. Noone, M. Krapcho, et al., SEER Cancer Statistics Review, 1975–2012, Bethesda, National Cancer Institute.

[3] M.D. Althuis, J.M. Dozier, W.F. Anderson, S.S. Devesa, L.A. Brinton, Global trends in breast cancer incidence and mortality 1973–1997, Int. J. Epidemiol. 34 (2) (2005) 405–412.

[4] J. Tang, R.M. Rangayyan, J. Xu, I. El Naqa, Y. Yang, Computer-aided detection and diagnosis of breast cancer with mammography: recent advances, IEEE Trans. Inf. Technol. Biomed. 13 (2) (2009) 236–251.

[5] M.L. Giger, H.-P. Chan, J. Boone, Anniversary paper: history and status of CAD and quantitative image analysis: the role of medical physics and AAPM, Med. Phys. 35 (12) (2008) 5799–5820.

[6] T. Uematsu, M. Kasami, S. Yuen, Neoadjuvant chemotherapy for breast cancer: correlation between the baseline MR imaging findings and responses to therapy, Eur. Radiol. 20 (10) (2010) 2315–2322.

[7] N. Biglia, V. Bounous, L. Martincich, E. Panuccio, V. Liberale, L. Ottino, R. Ponzone, P. Sismondi, Role of MRI (magnetic resonance imaging) versus conventional imaging for breast cancer presurgical staging in young women or with dense breast, Eur. J. Surg. Oncol. (EJSO) 37 (3) (2011) 199–204.

[8] M. Kriege, C.T. Brekelmans, C. Boetes, P.E. Besnard, H.M. Zonderland, I.M. Obdeijn, R.A. Manoliu, T. Kok, H. Peterse, M.M. Tilanus-Linthorst, et al., Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition, N. Engl. J. Med. 351 (5) (2004) 427–437.

[9] X. Liu, J. Tang, Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method, IEEE Syst. J. 8 (3) (2014) 910–920.

[10] R. Chaieb, A. Bacha, K. Kalti, F. Ben Lamine, Image features extraction for masses classification in mammograms, in: Proceedings of the 6th International Conference of IEEE Soft Computing and Pattern Recognition (SoCPaR), 2014, pp. 203–208.

[11] J.P. Thiran, B. Macq, Morphological feature extraction for the classification of digital images of cancerous tissues, IEEE Trans. Biomed. Eng. 43 (10) (1996) 1011–1020.

[12] D.-R. Chen, R.-F. Chang, Y.-L. Huang, Y.-H. Chou, C.-M. Tiu, P.-P. Tsai, Texture analysis of breast tumors on sonograms, in: Seminars in Ultrasound, CT and MRI, Vol. 21, Elsevier, 2000, pp. 308–316.

[13] I. Kitanovski, B. Jankulovski, I. Dimitrovski, S. Loskovska, Comparison of feature extraction algorithms for mammography images, in: Proceedings of the 4th International Congress on IEEE Image and Signal Processing (CISP), vol. 2, 2011, pp. 888–892.

[14] M.M. Eltoukhy, I. Faye, B.B. Samir, A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation, Comput. Biol. Med. 42 (1) (2012) 123–128.

[15] Y. Chen, Y. Lan, H. Ren, A feature selection method base on ga for cbir mammography cad, in: Proceedings of the 4th International Conference on IEEE Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 2, 2012, pp. 175–178.

[16] J. Chu, H. Min, L. Liu, W. Lu, A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation, Med. Phys. 42 (7) (2015) 3859–3869.

[17] P. Kang, S. Cho, EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems, Neural Information Processing, Springer, 2006, pp. 837–846.

[18] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, RUSBoost: a hybrid approach to alleviating class imbalance, IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. 40 (1) (2010) 185–197.

[19] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 42 (4) (2012) 463–484.

[20] K. Kira, L. A. Rendell, A practical approach to feature selection, in: Proceedings of the ninth international workshop on Machine learning, 1992, pp. 249–256.

[21] M. Abdel-Nasser, H.A. Rashwan, D. Puig, A. Moreno, Analysis of tissue abnormality and breast density in mammographic images using a uniform local directional pattern, Expert Syst. Appl. 42 (24) (2015) 9499–9511.

[22] D. ping Tian, A review on image feature extraction and representation techniques, Int. J. Multimed. Ubiquitous Eng. 8 (4) (2013) 385–396.

[23] N.R. Mudigonda, R.M. Rangayyan, J.L. Desautels, Gradient and texture analysis for the classification of mammographic masses, IEEE Trans. Med. Imaging 19 (10) (2000) 1032–1043.

[24] R.M. Haralick, K. Shanmugam, I.H. Dinstein, Textural features for image classification, IEEE Trans. Syst., Man Cybern. 6 (1973) 610–621.

[25] D. Cascio, F. Fauci, R. Magro, G. Raso, R. Bellotti, F. de Carlo, S. Tangaro, G. de Nunzio, M. Quarta, G. Forni, et al., Mammogram segmentation by contour searching and mass lesions classification with neural network, IEEE Trans. Nucl. Sci. 53 (5) (2006) 2827–2833.

[26] C. Gonzalez Rafael, E. Woods Richard, Digital Image Processing Third Edition Prentice-Hall.

[27] R.W. Conners, C.A. Harlow, A theoretical comparison of texture algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 3 (1980) 204–222.

[28] M.M. Galloway, Texture analysis using gray level run lengths, Comput. Graph. Image Process. 4 (2) (1975) 172–179.

[29] J. Hong, Gray level-gradient cooccurrence matrix texture analysis method, Acta Autom. Sin. 10 (1) (1984) 22–25.

[30] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.

[31] J. Kittler, Feature selection and extraction, Handbook of pattern recognition and image processing (1986) pp. 59–83.

[32] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. 4 (1) (1992) 1–58.

[33] P.-N. Tan, M. Steinbach, V. Kumar, et al., Introduction to data mining, Vol. 1, Pearson Addison Wesley Boston, 2006.

[34] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: Computational learning theory, Springer, 1995, pp. 23–37.

[35] Z. Zhou, Machine Learning, Tsinghua University Press, 2016.

[36] J.R. Quinlan, C4. 5: Programs for Machine Learning, Elsevier, 2014.

[37] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.

[38] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, S.Y. Philip, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (1) (2008) 1–37.

[39] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.

[40] T. Fawcett, ROC graphs: notes and practical considerations for researchers: notes and practical considerations for researchers, Mach. Learn. 31 (1) (2004) 1–38.

[41] D.M. Renz, J. Böttcher, F. Diekmann, A. Poellinger, M.H. Maurer, A. Pfeil, F. Streitparth, F. Collettini, U. Bick, B. Hamm, et al., Detection and classification of contrast-enhancing masses by a fully automatic computer-assisted diagnosis system for breast MRI, J. Magn. Reson. Imaging 35 (5) (2012) 1077–1088.

[42] K. Nie, J.-H. Chen, J.Y. Hon, Y. Chu, O. Nalcioglu, M.-Y. Su, Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI, Acad. Radiol. 15 (12) (2008) 1513–1525.

[43] E. Honda, R. Nakayama, H. Koyama, A. Yamashita, Computer-aided diagnosis scheme for distinguishing between benign and malignant masses in breast DCE-MRI, J. Digit. Imaging (2015) 1–6.

[44] S. Song, B.K. Seo, K.R. Cho, O.H. Woo, G.S. Son, C. Kim, S.B. Cho, S.-S. Kwon, Computer-aided detection (CAD) system for breast MRI in assessment of local tumor extent, nodal status, and multifocality of invasive breast cancers: preliminary study, Cancer Imaging 15 (1) (2015) 1.

Wei Lu received his B.Eng. degree in Electronic Engineering, and Ph.D. degree in signal and information processing from Tianjin University, Tianjin, China, in 1998 and 2003 respectively. He is currently an associate professor in the School of Electronic Information Engineering, Tianjin University. His teaching and research interests include digital filter design, digital multimedia technology, embedded system design, Web application design, and pattern recognition. He is now a senior member of the Chinese Institute of Electronics.

Zhe Li received his B.Eng. degree in Information and Communication Engineering from Zhejiang University, Hangzhou, China in 2014. He is now a master candidate in the School of Electronic Information Engineering in Tianjin University, Tianjin, China. His research involves machine learning, data mining, pattern recognition, and digital image processing.

Jinghui Chu received the B.Eng. degree in radio technology, and M.Eng. and Ph.D. degrees in signal and information processing all from Tianjin University, Tianjin, China, in 1991, 1997, and 2006 respectively.

She is currently an associate professor in the School of Electronic Information Engineering, Tianjin University. Her teaching and research interests include digital video technology and pattern recognition.