Machine Vision and Applications (2013) 24:1405–1420
DOI 10.1007/s00138-012-0459-8

SPECIAL ISSUE PAPER

Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles

Yungang Zhang · Bailing Zhang · Frans Coenen · Wenjin Lu

Received: 30 December 2011 / Revised: 22 June 2012 / Accepted: 17 September 2012 / Published online: 12 October 2012

Abstract Accurate and reliable classification of microscopic biopsy images is an important issue in computer assisted breast cancer diagnosis. In this paper, a new cascade Random Subspace ensembles scheme with reject options is proposed for microscopic biopsy image classification. The classification system is built as a serial fusion of two different Random Subspace classifier ensembles with rejection options to enhance the classification reliability. The first ensemble consists of a set of Support Vector Machine classifiers that converts the original K-class classification problem into a number of K 2-class problems. The second ensemble consists of a Multi-Layer Perceptron ensemble, that focuses on the rejected samples from the first ensemble. For both of the ensembles, the reject option is implemented by relating the consensus degree from majority voting to a confidence measure, and abstaining to classify ambiguous samples if the consensus degree is lower than some threshold. We also investigated the effectiveness of a feature description approach by combining Local Binary Pattern (LBP) texture analysis, statistics derived using the Gray Level Co-occurrence Matrix (GLCM) and the Curvelet Transform. While the LBP analysis efficiently describes local texture properties and the GLCM reflects global texture statistics, the Curvelet Transform is particularly appropriate for the representation of piece-wise smooth images with rich edge information. The combined feature description thus provides a comprehensive biopsy image characterization by taking advantages of their complementary strengths. Using a benchmark microscopic biopsy image dataset, obtained from the Israel Institute of Technology, a high classification accuracy of 99.25 % was obtained (with a rejection rate of 1.94 %) using the proposed system.

Keywords Breast cancer diagnosis · Biopsy image · Random subspace ensemble · Reject option · Combined feature

The project is funded by China Jiangsu Provincial Natural Science Foundation Intelligent Bioimages Analysis, Retrieval and Management (BK2009146).

Y. Zhang (✉) · F. Coenen
Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK
e-mail: yungang.zhang@liverpool.ac.uk

F. Coenen
e-mail: coenen@liverpool.ac.uk

B. Zhang · W. Lu
Department of Computer Science, Xi'an JiaoTong-Liverpool University, Suzhou 215123, People's Republic of China
e-mail: bailing.zhang@xjtlu.edu.cn

W. Lu
e-mail: wenjin.lu@xjtlu.edu.cn

Y. Zhang
School of Information Science, Yunnan Normal University, Kunming 650092, People's Republic of China

1 Introduction

Breast cancer accounts for nearly 1 in 4 cancers diagnosed in US women, it is also the most common type of cancer in women and the fifth most common cause of cancer death worldwide [1]. There is substantial evidence that there is a worldwide increase in the occurrences of breast cancer, especially in Asia, for example, China, India and Malaysia have recently experienced rapid increase in breast cancer incidence rates [2]. A recent study predicted that the cumulative incidence of breast cancer will increase to at least 2.2 million new cases among women across China over the 20-year period from 2001 to 2021 [3].

The most noticeable symptom of breast cancer is typically a lump or a tumor that feels different from the rest of the breast tissue. However, it is not easy to distinguish a malignant tumor from a benign one because there are structural similarities between the two. To accurately identify the structural differences, physicians have to cautiously study a patient's clinical history and make various medical examinations supported by imaging using mammography or ultrasound. However, the precise diagnosis of a breast tumor can only be obtained through some form of biopsy where by a small sample of cells or tissue is removed for examination. Typical biopsy processes for breast cancer analysis include Fine-Needle Aspiration (FNA), core needle, and excisional biopsy [4]. Among these FNA is the most convenient because it involves the use of very small needles (smaller than those used for blood tests) [5]. This deterministic diagnosis is vital as the potency of the cytotoxic drugs administered during treatment can be life threatening.

As there is always a subjective element related to the pathological examination of a biopsy, an automated technique will provide valuable assistance for physicians. Recent years have witnessed a large increase in research related to computer assisted breast cancer diagnosis. A large focus with respect to biopsy image analysis has been on automated cancer type classification. Many recent studies have revealed that biopsy images can be properly classified, without requiring perfect segmentation if suitable image feature descriptions are chosen [6–8]. Tabesh et al. aggregated color, texture, and morphometric cues at the global and histological object levels for classification, achieving 96.7 % classification accuracy in classifying tumor and non-tumor images [9]. The wavelet package transform coupled with local binary patterns were used for meningioma subtype classification in [10]. This research, and similar work, demonstrated that by combining different image description features it is possible to improve medical image classification performance.

A great number of machine learning methods have been proposed to design accurate classification systems for various medical images [11]. Among them, ensemble learning has attracted much attention due to the good performance from many applications in medicine and biology [12]. Ensemble learning is concerned with mechanisms to combine the results of a number of weak learning systems. A weak learner is defined to be a classifier which is only slightly correlated with the true classification, it can label examples better than random guessing. In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification [13]. In the case of ensemble classification, ensemble learning is concerned with the integration of the results of a number of classifiers (often called as 'base classifiers') [14] to develop a strong classifier with good generalization performance, therefore, 'base classifiers' are also referred as 'weak classifiers'.

Among the representatives of ensemble learning, the Random Subspace (RS) method [15] is often quoted as an efficient way of combining the results of a number of classifiers. A recent application of RS for functional Magnetic Resonance Imaging (fMRI) classification has shown promising results [16]; here RS outperformed single classifiers as well as some of the most widely used alternative classifier ensemble techniques such as bagging, AdaBoost, random forests and rotation forests. The same outcome has also been reported in the context of RS ensemble based gene expression classification [17]. RS divides the input feature space into subspaces; each subspaces is formed by randomly picking features from the entire space, features may be repeated across subspaces.

In previous studies of medical images classification, accuracy was the only objective; the aim was to produce a classifier that featured the smallest error rate possible. In many applications, however, it is more important to address the reliability issue in classifier design by introducing a reject option which allowed for an expression of doubt. The objective of the reject option is thus to improve classification reliability by leaving the classification of "difficult" cases to human experts. Since the consequences of misclassification may often be severe when considering medical image classification, clinical expertise is desirable so as to exert control over the accuracy of the classifier in order to make reliable determinations.

Classification with a rejection option has been a topic of interest in pattern recognition. Multi-stage classifiers are ensembles where individual classifiers have a reject option [18]. Cascading [19] is a scheme to support multi-stage classification. At the first stage of a cascading system, the system constructs a simple rule using a properly generalized classifier; based on its confidence criterion, it is likely that the rule will not cover some part of the space with sufficient confidence. Therefore, at the next stage, cascading builds a more complex rule to focus on those uncovered patterns. Eventually there will remain few patterns which are not covered by any of the prior rules, these patterns can then be dealt with using an instance-based nonparametric technique which is good at unrelated, singular points [20]. Many cascading multi-stage classifier architectures have been proposed [21–24] and plenty of promising results have been achieved in medical and biological classification applications, such as microarray data classification [25] and gene expression data classification [26].

In this paper, we propose and evaluate a novel cascade scheme, comprised of two random subspace ensembles, to be applied to microscopic biopsy image classification with a reject option. The first stage of our cascade scheme consists of an ensemble of SVMs with reject option to classify patterns with high level of confidence. The more complex and slower second stage, which is an ensemble of MLPs,

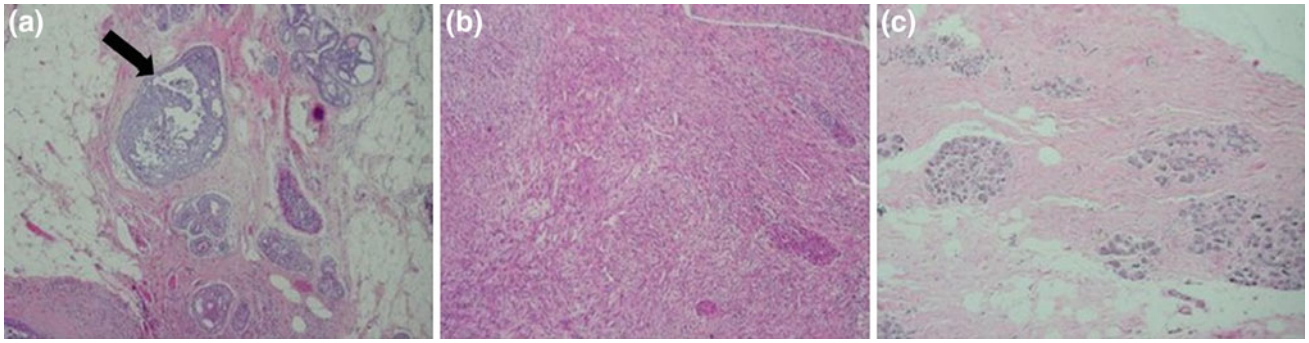

Fig. 1 Typical image instances. **a** carcinoma in situ: tumor confined to a well-defined small region, usually a duct (*arrow*); **b** invasive: breast tissue completely replaced by the tumor; **c** normal: normal breast tissue, with ducts and finer structures

deals with the rejected patterns from stage 1, and is designed to make further classifications or rejections. Compared with some earlier cascading classifier paradigms, our proposed system is composed of two different ensembles. In the first stage, a one-vs-all SVM ensemble is employed to classify "straight forward" samples (thus obtaining high accuracy) and reject those which are less straight forward or ambiguous. Only samples for which the ensemble's confidence score, in terms of consensus degree, is greater than a certain threshold will be classified. The second stage consists of a random subspace ensemble of MLPs which operates using majority voting, any samples that have a low consensus degree will be rejected for further consideration by human experts. It is suggested that classification with the proposed cascaded ensembles will provide an efficient means to simultaneously reduce the error rate and enhance the reliability by controlling the accuracy-rejection trade-off.

The rest of this paper is organized as follows: the breast cancer biopsy image dataset used in our work and the image feature extraction methods are introduced in Sect. 2. In Sect. 3, we described and theoretically analyzed the proposed two-stage ensemble cascading system in detail. In Sect. 4, the experimental results are given based on the adopted benchmark image dataset. We compared the proposed cascading system with its component classifiers as well as some widely used aggregation techniques, such as bagging and AdaBoost. The paper ends with some conclusions in Sect. 5.

2 Biopsy images and feature descriptions

In this section we will first introduce, in Sect. 2.1, the benchmark breast cancer biopsy image dataset. The proposed image feature extraction methods are then introduced in Sect. 2.2. The choice of features for describing the initial biopsy images depends on the nature of the input images. For biopsy image classification, calculating global features to estimate the global appearance of the images is an effective approach. In this work, we propose to use three image descriptors for biopsy image feature extraction: (1) local binary patterns (Sect. 2.2.1), (2) gray level co-occurrence matrixes (Sect. 2.2.2) and (3) the curvelet transform (Sect. 2.2.3).

2.1 Breast cancer biopsy images

With respect to the work described in this paper a breast cancer benchmark biopsy images dataset from the Israel Institute of Technology[1] was used. The image set consists of 361 samples, of which 119 were classified by a pathologist as normal tissue, 102 as carcinoma in situ, and 140 as invasive ductal or lobular carcinoma. The samples were generated from breast tissue biopsy slides, stained with hematoxylin and eosin. They were photographed using a Nikon Coolpix® 995 attached to a Nikon Eclipse® E600 at magnification of $\times 40$ to produce images with resolution of about 5 u per pixel. No calibration was made, and the camera was set to automatic exposure. The images were cropped to a region of interest of 760×570 pixels and compressed using the lossy JPEG compression. The resulting images were again inspected by a pathologist to ensure that their quality was sufficient for diagnosis. Figure 1 presents three sample images of healthy tissue, tumor in situ and invasive carcinoma.

2.2 Feature descriptions

Shape feature and texture feature are critical factors for distinguishing one image from another. For the biopsy image discrimination, shapes and textures are also quite effective. As we can see from Fig. 1, three kinds of biopsy images have visible differences in cell externality and texture distribution. Thus, we use Local Binary Patterns (LBPs) for extracting local textural features, Gray Level Co-occurrence Matrix (GLCM) statistics for representing global textures and the Curvelet Transform for shape description.

[1] http://ftp.cs.technion.ac.il/pub/projects/medic-image.

2.2.1 Local binary patterns

Local Binary Patterns (LBPs) were first introduced as a texture descriptor for summarizing local gray-level structures [27,28], LBPs are generated by taking a local neighborhood around each pixel into account, thresholding the pixels of the neighborhood at the value of the central pixel and then using the resulting binary-valued image patch as a local image descriptor. In other words, a binary code of 0 or 1 is assigned to each neighborhoods pixel. The binary code of each pixel in the case of a 3×3 neighborhoods would form an 8 bits code. In this manner, a single scan through an image can generate LBP codes for each pixel.

Formally, the LBP operator takes the form

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \quad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{1}$$

where g_c is the gray value of the central pixel, g_p is the value of its neighbors, P is the total number of neighbors and R is the radius of the neighborhood.

A useful extension to the original LBP operator is the so-called uniform patterns [27]. An LBP is "uniform" if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. For example, 11100001 (with two transitions) is a uniform pattern, whereas 11110101 (with four transitions) is a non-uniform pattern. The uniform LBP describes those structures which contain at most two bitwise (0 to 1 or 1 to 0) transitions. Uniformity represents important structural features such as edges, spots and corners. Ojala et al. [27] observed that although only 58 of the 256 eight-bit patterns are uniform, nearly 90 % of all observed image neighborhoods are uniform. We use the notation $LBP^u_{P,R}$ for the uniform LBP operator, meaning a neighborhood of P sampling points on a circle of radius R. The superscript u stands for using uniform patterns and labeling all remaining patterns with a single label. The number of labels for a neighborhood of 8 pixels is 256 for standard LBP and 59 for $LBP^u_{8,1}$.

A common practice when applying an LBP coding over an image is to generate a histogram of the labels, where a 256-bin histogram represents the texture description of the image and each bin can be regarded as a micro-pattern. The distribution of these patterns represents the whole structure of the texture. The number of patterns in an LBP histogram can be reduced by only using uniform patterns without losing much information. As noted above, there are 58 different uniform patterns in an 8-bit LBP representation, the remaining patterns can be assigned in one non-uniform binary number, thus representing the texture structure with a 59-bin histogram instead of using 256 bins.

LBP has been shown to be an efficient image texture descriptor. Recently, a complete modeling of the local binary pattern operator was proposed and the associated Complete LBP (CLBP) scheme developed for texture classification [28]. Different to traditional LBP, in CLBP, a local region is represented by its center pixel and a Local Difference Sign-Magnitude Transform (LDSMT). With a global thresholding, the center pixel is coded by a binary code and the binary map is called $CLBP\_C$. Two other complementary components are also obtained by LDSMT: the difference signs and the difference magnitudes, two operators $CLBP\_S$ and $CLBP\_M$ are used to code them. The framework of CLBP is presented in Fig. 2. The CLBP could achieve much better rotation invariant texture classification results than conventional LBP based schemes.

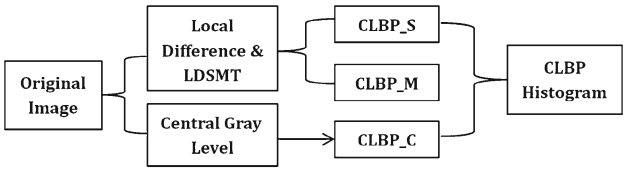

Fig. 2 Framework of CLBP

We briefly review three operators in CLBP here, namely, $CLBP\_S, CLBP\_M$ and $CLBP\_C$. Given a central pixel g_c and its P neighbors g_p, $p = 0, 1, \ldots, P-1$, the difference between g_c and g_p can be calculated as $d_p = g_p - g_c$. The local difference vector $[d_0, \ldots, d_{P-1}]$ describes the image local structure at g_c, d_p can be further decomposed into two components:

$$d_p = s_p * m_p, \quad \text{and} \quad \begin{cases} s_p = \text{sign}(d_p) \\ m_p = |d_p| \end{cases} \tag{2}$$

where $s_p = 1$, when $d_p \geq 0$, otherwise, $s_p = 0$. m_p is the magnitude of d_p. Equation (2) is called the local difference sign-magnitude transform (LDSMT).

The $CLBP\_S$ operator is defined as the original LBP operator in Eq. (1).

The $CLBP\_M$ operator is defined as:

$$CLBP\_M_{P,R} = \sum_{p=0}^{P-1} t(m_p, c)2^p,$$

$$t(x, c) = \begin{cases} 1 & \text{if } x \geq c \\ 0 & \text{if } x < c \end{cases} \tag{3}$$

where c is a threshold set as the mean value of m_p from the whole image.

The $CLBP\_C$ operator is coded as:

$$CLBP\_C_{P,R} = t(g_c, c_I) \tag{4}$$

where t is defined in Eq. (3) and c_I is a threshold set as the average gray level of the whole image.

In this work, we use the 3D joint histogram of these three operators to generate textural features of breast cancer biopsy images, according to [28], the joint combination of

the three components gives better classification than conventional LBP and provides a smaller feature dimension.

2.2.2 Statistics from gray level co-occurrence matrix

Global texture distribution is one of the important characteristics used in identifying objects or regions of interest in an image. The co-occurrence probabilities provide a second-order method for generating texture features [29]. The basis for features used here is the gray level co-occurrence matrix, the matrix is square with dimension N_g, where N_g is the number of gray levels in the image. Element $[i, j]$ of the matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j [30], the matrix can be seen in Eq. (5).

$$\mathbf{C} = \begin{bmatrix} p(1,1) & p(1,2) & \cdots & p(1,N_g) \\ p(2,1) & p(2,2) & \cdots & p(2,N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g,1) & p(N_g,2) & \cdots & p(N_g,N_g) \end{bmatrix} \tag{5}$$

With respect to the work described in this paper, a total of 22 features were extracted from gray level co-occurrence matrices in our work, these are listed in Table 1. Each of these statistics has a qualitative meaning with respect to the structure within the GLCM, for example, dissimilarity and contrast measure the degree of texture smoothness, uniformity and entropy reflect the degree of repetition amongst the gray-level pairs, and correlation describes the correlation between the gray-level pairs. For details of these statistical features, see [29–32].

Table 1 Features extracted from gray level co-occurrence matrix

Index	Features	Index	Features
1	Energy	12	Sum of squares
2	Entropy	13	Sum average
3	Dissimilarity	14	Sum variance
4	Contrast	15	Sum entropy
5	Inverse difference	16	Difference variance
6	Correlation	17	Difference entropy
7	Homogeneity	18	Information measure of correlation (1)
8	Autocorrelation	19	Information measure of correlation (2)
9	Cluster shade	20	Maximal correlation coefficient
10	Cluster prominence	21	Inverse difference normalized
11	Maximum probability	22	Inverse difference moment normalized

2.2.3 Curvelet transform

The Curvelet transform [33–37] is one of the latest developments in non-adaptive transforms. Compared to the wavelet transform, the curvelet transform provides a more sparse representation of an image, with improved directional elements and better ability to represent edges and other singularities along curves. Sparse representation usually offers better performance with its capacity for efficient signal modeling. So far, successful applications of the curvelet transform have been found in many medical and biological image analysis tasks, including digital mammogram analysis [38] and phenotype recognition [39].

In the curvelet transform, fine-scale basis functions are long ridges; the shape of the basis functions at scale j is 2^{-j} by $2^{-j/2}$ so the fine-scale bases are skinny ridges with a precisely determined orientation. The curvelet coefficients can be expressed by:

$$c(j,l,k) := \langle f, \varphi_{j,l,k} \rangle = \int_{R^2} f(x)\varphi_{j,l,k}(x)\mathrm{d}x \tag{6}$$

where $\varphi_{j,l,k}$ denotes the curvelet function, and j, l and k are the variables of scale, orientation, and position, respectively.

In the last few years, several discrete curvelet transforms have been proposed. The most influential approach is based on the Fast Fourier Transform (FFT) [36]. In the frequency domain, the curvelet transform can be implemented with φ by means of the window function U. Defining a radial window $W(r)$ and an angular window $V(t)$ as follows:

$$\sum_{j=-\infty}^{\infty} W^2(2^j r) = 1, \qquad r \in (3/4, 3/2) \tag{7}$$

$$\sum_{j=-\infty}^{\infty} V^2(t-1) = 1, \qquad t \in (-1/2, 1/2) \tag{8}$$

where W is a frequency domain variable and r and θ are polar coordinates within the frequency domain. For each $j \geq j_0$, U_j is defined over the Fourier domain by:

$$U_j(r,\theta) = 2^{3j/4} w(2^{-j}r) v\left(\frac{2^{[j/2]}\theta}{2\pi}\right) \tag{9}$$

where $[j/2]$ denotes the integer part of $j/2$.

The fastest curvelet transform currently available is curvelets via wrapping [36], which will be used for our work. From the curvelet coefficients, some statistics can be calculated from each of these curvelet sub-bands. In this paper, the mean μ, the standard deviation δ and the entropy H are used as the simple features. If n curvelets are used for the transform, $3n$ features $G = [G_\mu, G_\delta, H]$ are obtained,

where $G_\mu = [\mu_1, \mu_2, \ldots, \mu_n]$, $G_\delta = [\delta_1, \delta_2, \ldots, \delta_n]$ and $H = [h_1, h_2, \ldots, h_n]$. A $3n$ dimensional feature vector can be used to represent each image in the dataset.

2.2.4 Combined features

Each feature extracted from the above three descriptors characterizes individual aspects of image content. The joint exploitation of different image descriptions is often necessary to provide a more comprehensive description in order to produce classifiers with higher accuracy. Using 5 levels of the curvelet transform, 82 sub-bands of curvelet coefficients are computed, therefore, a 246-dimensional curvelet feature vector is generated for each image. With a 64 gray-level quantization, we used 10 different relative interpixel distances to generate 10 different gray level co-occurrence matrices for each image. The 22 statistics listed in Table 1 are computed for each of these 10 gray level co-occurrence matrices, thus, we have a 220-dimensional GLCM feature vector for each image. The CLBP feature vector of each image has a dimension of 200. The three feature vectors are normalized, respectively, into the range of $[-1, 1]$, then concatenated together to produce a 666-dimensional feature vector of each image for classification. One of the difficulties of multiple feature aggregation lies in the high dimensionalities of the feature space. However, by using Random Subspace classifier ensembles (see following section) this problem can be resolved due to its dimension reduction capability.

Due to the differences existing in different molecular imaging devices and staining methods, histology images of biopsies may change significantly in colors and intensities. The above explained feature extractors can cope with this situation effectively, since all of them work in the grayscale color space. Before feature extraction, all the biopsy images will be converted from a chromatic color space to a grayscale color space with an intensity interval from 0 to 255. The conversion eliminates the adverse effects from color and intensity variations because the feature extractors work in the same space. Moreover, as the features are extracted from the whole images, the distribution and structures of different patterns such as tissues, ducts, fat and tumors, will be automatically described by the feature extractors, thus, they will not affect the performance of the combined feature.

3 Serial fusion of random subspace ensembles

Although many supervised learning algorithms such as neural networks, the k-nearest neighbor algorithm and SVM have been extensively applied to many medical image classification problems, few of them has addressed the issue of classification reliability (the extent that one can rely upon a given prediction). Note that we are interested in the assessment of a classifier's performance on a single example such as the diagnosis associated with am individual patient. In such cases an overall quality measure of a classifier (e.g., classification accuracy) would not provide the desired information, even where good accuracies are achieved using some state-of-art method. With respect to some real applications, such as medical diagnosis, highly reliable classifiers are required so that a correct therapeutic strategy can be selected. Therefore, it is desirable to have a reject option in order to avoid making a wrong decision when classifier is presented with ambiguous input, i.e., an option to withhold a classifier decision.

In this paper a new two-stage classifier for the breast cancer biopsy image classification, consisting of a random subspace ensembles with reject option, is proposed. With respect to the work described in this paper, we adopted the definitions of recognition rate, rejection rate and reliability proposed in [40], as presented below, so as to facilitate the performance evaluation of classifiers with a reject option:

- Recognition rate (RR) = no. of correctly recognized images/no. of testing images
- Rejection rate (ReR) = no. of rejected images/no. of testing images.
- Reliability (RE) = (no. of correctly recognized images + no. of rejected images)/no. of testing images.
- Error rate (ER) := 100 % − reliability.

From the above we can see that Reliability = Recognition rate + Rejection rate. According to this definition of reliability, high reliability can be achieved with an appropriate trade-off between error rate and rejection rate.

3.1 Reject option for classification

The optimal classification rule with reject option was defined by Chow [21]. Consider a binary classification task with an instance dataset $X = \{x_1, x_2, \ldots, x_m\}$ and a class label set $C = \{-1, 1, 0\}$ where class 0 is the reject option. We need to seek a classification rule, $L\ (X \Rightarrow C)$such that $L(x) = 0$ indicates that no definite judgement will be made for x and a reject option taken. Chow's rule rejects a pattern if the maximum of its a posterior probabilities is lower than a predefined threshold t, the pursuit of maximum of the posterior probabilities can be identified as a measure of classification reliability. Such a rule can be expressed as:

$$f(x) = \begin{cases} \text{argmax}_{C_i}(p(C_i|x)) & \text{if } \max_{C_i}\ (p(C_i|x)) \geq t \\ \text{reject} & \text{if } \forall_i\ p(C_i|x) < t \end{cases} \tag{10}$$

where $p(C_i|x)$ is the posterior probability, which can be obtained by Bayes formula.

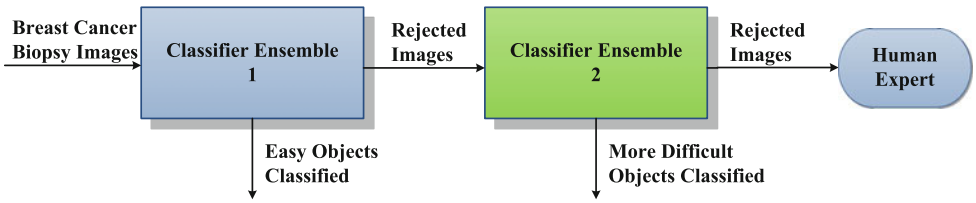

Fig. 3 Operation of the hybrid classification scheme comprising a cascade of two Random Subspace classifier ensembles

The rejection rate is the probability that the classifier rejects a given example:

$$p(reject) = \int_{\text{reject}} p(x)dx = p(\max(p(C_i|x)) < t). \qquad (11)$$

In Chow's theory, an optimal classifier can be found only if the true posterior probabilities are known. This is rarely reachable in real applications.

The key issue with respect to the reject option is to define the threshold t, in our work, we do not deeply consider the optimal error-reject trade-off. We used different rejection thresholds and the results of rejection against accuracies and reliabilities were compared.

3.2 A cascade two-stage classification scheme

As already noted, it has been demonstrated that classification accuracy can be enhanced by using an ensemble of classifiers. Over the last few years a number of successful ensemble methods have been proposed [14,16]. The most popular method for creating a classifier ensemble is to build multiple parallel classifiers, and then to combine their outputs according to some fusion strategy. Alternatively, a serial architecture can be adopted with different classifiers arranged in cascade form such that the output of a classifier acts as the input to another classifier. In this paper, we will propose a hybrid classification scheme which serially connects two parallel random subspace ensemble classifiers (Fig. 3). Note that all classifiers have a rejection option.

In our current implementation the first ensemble (Classifier Ensemble 1 in Fig. 3) consists of a collection of SVM classifiers, the second (Classifier Ensemble 2 in Fig. 3) consists of a collection of MLP classifiers. From Fig. 3 it can be seen that rejected samples from Classifier Ensemble 1 are passed to Ensemble 2, any samples that remain rejected once Classifier Ensemble 2 has been applied are passed to a human expert for "adjudication".

SVM and MLP have obtained better performance than other kinds of classifiers in many medical image analysis tasks, especially in histopathological image analysis [11], therefore, they have been selected as the base classifiers in our two ensembles. The proposed cascade system here is consistent with a principle in statistical pattern recognition that an improved classification performance can be expected when a local classifier is appended after a global one [41]. The SVM ensemble in the first stage is trained as a global classifier. Compare with SVM, the MLP is relatively local, since it has been proven that a feed-forward network of just two layers (not including the input layer) is enough to approximate any continuous function [42]. Note that the classification performance of the whole system will not change too much if we use another SVM ensemble in the second stage, because under the same training strategy, the obtained support vectors in stage 1 and stage 2 will be very similar.

Another reason we use different base classifiers for the two ensembles is to achieve'diversity' between classifiers, which is also deemed as an important factor for the success of ensemble learning [43]. Making use of different individual classifiers in an ensemble can improve the ensemble performance, here we expand the concept to employ different base classifiers for the two ensembles to improve the 'diversity' between the ensembles.

The major issue for designing the above grid classification system is to decide when a pattern is covered by a rule and should be classified accordingly, and when it should be rejected and either passed on to the second ensemble or the human expert (depending on which stage in the process we are at). The reject option has been formalized in the context of statistical pattern recognition according to the minimum risk theory presented in [21] and [44]. Intuitively, a suggested classification should be rejected if the confidence in that classification is below some threshold.

The standard approach to rejection in classification is to estimate the class posteriors, and to reject classifications that have a low class posterior probabilities [21]. To simplify the design of the SVMs in the first ensemble with appropriate posteriors estimation, we decompose the multi-label classification problems with K classes ($K = 3$ in current work) into K independent two-class problems (the one-vs-all approach where each classifier classifies records as belonging or not belonging to a class). The desired multi-class classification can then be conducted according to the output of the binary classifiers.

To estimate class posteriors from SVM's outputs, a mapping can be implemented using the following sigmoid function [45]:

$$P(y = +1|\mathbf{x}) = \frac{1}{1 + \exp(a\rho(\mathbf{x}) + b)} \qquad (12)$$

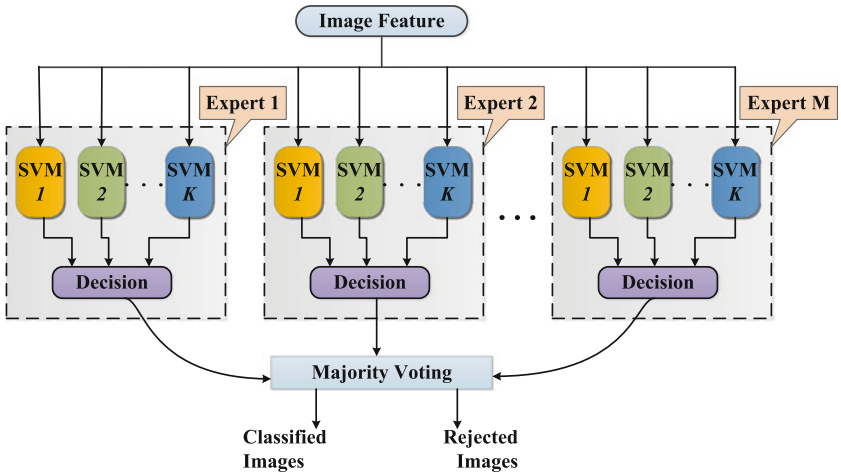

Fig. 4 SVM ensemble with rejection option in stage 1, which consists of a set of binary SVMs (experts)

where the class labels are denoted as $y = +1, -1$, while a and b are constant terms to be defined on the basis of sample data. Such a method provides estimates of the posterior probabilities that are monotonic functions of the output $\rho(x)$ of an SVM. This implies that Chow's rule applied to such estimates is equivalent to the rejection rule obtained by directly applying a reject threshold on the absolute value of the output $\rho(x)$ [23].

In our scheme, K binary SVM classifiers are constructed for K different image classes ($K = 3$). And we term such K collection of binary SVMs an expert to avoid the confusion with ensemble. The ith SVM output function P_i is trained taking the examples from i-th class as positive and the examples from all other classes as negative. In another word, each binary SVM classifier was trained to act as a class label detector, outputting a positive response if its label is present and a negative response otherwise. Therefore, for example, a binary SVM trained as a "in situ detector" would classify between in situ and not in situ. For a new sample x, the corresponding SVM assigns it to the class with the largest value of P_i following

$$Class = \text{argmax} \;\; P_i, \quad i = 1, \ldots, n \tag{13}$$

where P_i is the signed confidence measure of the ith SVM classifier. Such a SVM expert can then act as a base classifier in the stage 1 ensemble, trained with randomly chosen subsets of all available features (i.e., random subspace) following the Random Subspace strategy [15]. In the random subspace method, base classifiers are learned from random subspaces of the data feature space. In other words, the ensemble is trained by dividing the feature space randomly into subsets and submits each one to a base SVM expert.

As we aim to construct a serially fused, cascade classifier ensembles in order to produce a high confidence classification, it is essential to examine the output from the VM ensemble consisting of the base SVM experts. In combining the decisions from the M experts, a sample is assigned the class for which there is a predefined consensus degree, or when at least t of the experts are agreed on the label, otherwise, the sample is rejected, the threshold t can be decided in advance, for example, a simple rule as follows can be used to decide the value of t.

$$t \geq \begin{cases} \frac{M}{2} + 1 & \text{if } M \; is \text{ even} \\ \frac{M+1}{2} & \text{if } M \; is \text{ odd.} \end{cases} \tag{14}$$

Since there can be more than two classes, the combined decision is deemed to be correct when a majority of the experts are correct, but wrong when a majority of the decisions are wrong. Obviously, t is a tunable threshold that controls the rejection rate, and we use t to relate the consensus degree from the majority voting to the confidence measure, and abstain from classifying ambiguous samples. A rejection is considered neither correct nor wrong, so it is equivalent to a neutral position or an abstention [46]. Figure 4 further explains the principle of the SVM ensemble in stage 1.

The rejected samples from the SVM ensemble in stage 1 will be handled by the second ensemble, which is a Random Subspace ensemble of neural network classifiers, simultaneously trained with the stage 1 SVM ensemble. The neural network classifier is a Multiple Layer Perceptron (MLP), which has one hidden layer with a few hidden neurons and 3 output nodes, each representing a class label. The activation functions for the hidden and output nodes are a logistic sigmoid function and linear function, respectively. Following the principle of RS, a number of individual MLP models are trained on randomly chosen subsets of all available features. That is, an ensemble of MLP classifiers is created with each base classifier trained on an individual subspace by randomly selecting features from the entire space.

The last step of the second Random Subspace ensemble is to combine the base MLP models to give final decisions following the similar procedure of majority voting as in the first stage, as shown in Fig. 5. In combining the decisions from the M base MLPs, a sample (selected from the collection of

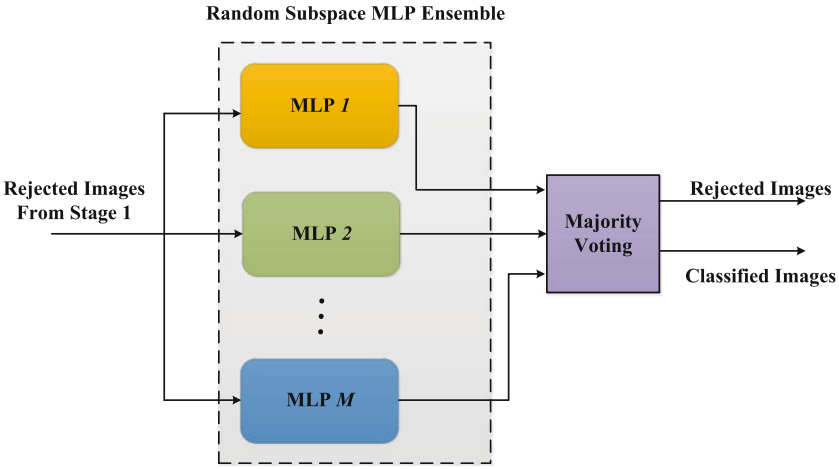

Fig. 5 Illustration of the stage 2 Random Subspace classifier ensemble which consists of a set of MLPs

rejected samples from stage 1) is assigned the class label when at least t of the MLPs are agreed on the decision. Otherwise, the sample is rejected. Again, t is the threshold that decide the rejection rate. The consensus degree from the ensemble acts as confidence measure to switch between acceptance and rejection.

3.3 Theoretical analysis of the ensemble cascade

If we have $p(C_i)$ as the prior probability of observing class C_i, the posterior probability of class C_i when given an instance vector x can be calculated as:

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)} = \frac{p(x|C_i)p(C_i)}{\sum_{i=1}^{N} p(x|C_i)p(C_i)} \tag{15}$$

where N is the number of classes, $p(x|C_i)$is the conditional probability of x given a class C_i, and $p(x)$ is the probability of x.

We adopted the mechanism proposed in [47] to derive the error rate of our system. For both stages in our scheme, given an input instance x, the proposed classification is accepted or rejected according to the highest posterior probability for all the classes: $\max_{j\in[1,\ldots,N]} p(C_j|x)$. Since the result of our classifiers is only an approximation of the real situation, we use $S_{i\ (i=1,\ldots,N)}$ to denote the approximation posterior probability for each class obtained by our system. Let $MAX_p^1 = \max_{j\in[1,\ldots,N]} p(C_j|x)$ denote the real posterior probabilities for all classes given an instance x, and $MAX_S^1 = \max_{i\in[1,\ldots,N]} S_i^1$ represents the approximation posterior probabilities obtained by stage 1 of our system. The error rate of stage 1 ϵ_1 can be obtained by:

$$\epsilon_1 = \int_A (1 - MAX_S^1) p(x)\mathrm{d}x \tag{16}$$

where A is the region composed of all accepted instances. Using some simple manipulations on Eq. (16), we then get the following:

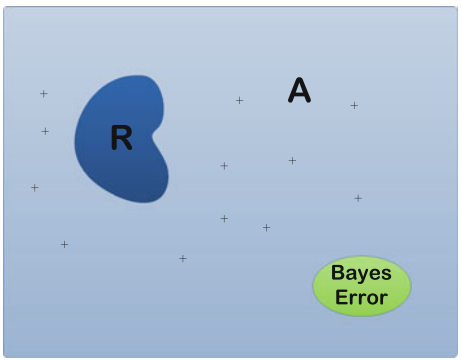

Fig. 6 Error rate of stage 1

$$\begin{aligned}
\epsilon_1 &= \int_A (1 - MAX_S^1) p(x)\mathrm{d}x \\
&= \int_A (1 - MAX_p^1 + MAX_p^1 - MAX_S^1) p(x)\mathrm{d}x \\
&= \int_A (1 - MAX_p^1) p(x)\mathrm{d}x \\
&\quad + \int_{A\cap I^S} (MAX_p^1 - MAX_S^1) p(x)\mathrm{d}x
\end{aligned}$$

where I^S is the region composed of all the instances that satisfy $MAX_p^1 - MAX_S^1 \neq 0$, which means that for some input instances, the results of our classifiers are different from the real ones. Notice that the first term of ϵ_1 is in fact the optimal Bayes error $\int(1 - p(x))p(x)\mathrm{d}x$. The second term comes from the errors generated during stage 1. This situation can be illustrated as in Fig. 6, where R represents the rejected patterns, A represents the patterns accepted by the classifier and the crosses represent erroneous classifications made by the ensemble of stage 1.

The same procedure can be used to analyze the error rate of stage 2. Instead of a wide input instance space, stage 2

only processes the rejected instances from stage 1. Let R denote the region composed by all the rejected instances from stage 1, $R = \{x|\max(p(C_i|x)) < t\}$, $MAX_p^2 = \max_{j\in[1,\ldots,N]} p(C_j|x)$ and $MAX_S^2 = \max_{i\in[1,\ldots,N]} S_i^2$. The error rate of stage 2 can then be obtained by:

$$\epsilon_2 = \int_R (1 - MAX_p^2)p(x)\mathrm{d}x + \int_{R\cap I^M} (MAX_p^2 - MAX_S^2)p(x)\mathrm{d}x \quad (17)$$

where $I^M = \{x|MAX_p^2 - MAX_S^2 \neq 0\}$, which represents the errors generated by the stage 2 ensemble.

Given the above, the error rate of the whole system can be calculated as:

$$\begin{aligned}\epsilon &= \epsilon_1 + \epsilon_2 \\ &= \int_A (1 - MAX_p^1)p(x)\mathrm{d}x + \int_R (1 - MAX_p^2)p(x)\mathrm{d}x \\ &\quad + \int_{A\cap I^S} (MAX_p^1 - MAX_S^1)p(x)\mathrm{d}x \\ &\quad + \int_{R\cap I^S} (MAX_p^2 - MAX_S^2)p(x)\mathrm{d}x \\ &= \epsilon_{\text{Bayes}} + \int_{A\cap I^S} (MAX_p^1 - MAX_S^1)p(x)\mathrm{d}x \\ &\quad + \int_{R\cap I^M} (MAX_p^2 - MAX_S^2)p(x)dx. \end{aligned} \quad (18)$$

From Eq. (18), for approaching the goal that $\epsilon = \epsilon_{\text{Bayes}}$, we must set $A \cap I^S = \emptyset$ and $R \cap I^M = \emptyset$. This means that even if both stages are not optimal, we still have chance to reach the optimal classification error rate. However, this can rarely be expected in real classification tasks.

Different from many existing cascade systems, we use classifier ensembles in our architecture. As has already been pointed out in [40], under the sum voting ensemble schemes, the variance of the ensemble is less than that of the individual classifier and a smaller variance in an ensemble will lead to a lower error rate than any individual classifier. From the above theoretical analysis, with a cascade system composed of two ensembles, a lower error rate can be expected than when using non-ensemble or non-cascade methods.

4 Experiments

MATLAB version 7 was used to implement the software in the current work. Six different individual classifiers were applied to the image dataset first, their results are compared and analyzed. Then several popular classifier ensemble methods were employed to construct the ensemble classifiers. In order to ascertain the effectiveness of the proposed feature combinations, several different feature combinations were computed and compared. The performance (accuracy and reliability) of the proposed two-stage ensemble cascade scheme was evaluated using different ensemble sizes and different rejection rates.

4.1 Comparison among single classifiers

In this section, we show the results obtained using six different classifiers on the biopsy image dataset where each image was described in terms of the three kinds of features introduced in Sect. 2. The six classifiers were (1) kNN, $k = 3$, (2) single MLP, (3) single SVM, (4) Logistic Regression [48], (5) Fisher Linear Discrimination [48] and (6) Naive Bayesian Classifier [49]. For MLP, we experimented with a three-layer network. Specifically, the number of inputs is the same as the number of features, one hidden layer with 20 units was used and a single linear unit representing the class label. The network was trained using the Conjugate Gradient learning algorithm for 500 epochs. The library for support vector machines, LIBSVM,[2] was used for the experiments. We used the radial basis function kernel for the SVM classifier. The parameter γ that defines the spread of the radial function was set to 5.0 and the parameter C that defines the trade-off between the classifier accuracy and the margin to 3.0. For the microscopic biopsy images, we randomly split it into training and testing sets, each time with 20 % of each class' images reserved for testing while the rest was used for training. The classification results were then averaged over 100 runs, such that each run used a random split of the data for the training and testing sets.

In Fig. 7, we compared the classification accuracies with respect to the six classifiers. The averaged classification accuracies of the MLP and SVM were 94.90 and 94.85 %, respectively, which are far beyond the other four classifiers. The standard deviations of the classification accuracies are also compared in Fig. 7. Although the FLD has the smallest averaged standard deviation (0.0571) on its classification accuracy, it has the lowest classification performance. The averaged standard deviations of MLP and SVM are 0.0934 and 0.1040, respectively, which are relatively smaller than that of the other classifiers, which means they are more stable with respect to classification performance.

Figure 8 presents a box plot of the classification results obtained by these six single classifiers on the biopsy image dataset. From the figure it can be seen that the MLP and SVM classifiers have small variance ranges in classification

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm.

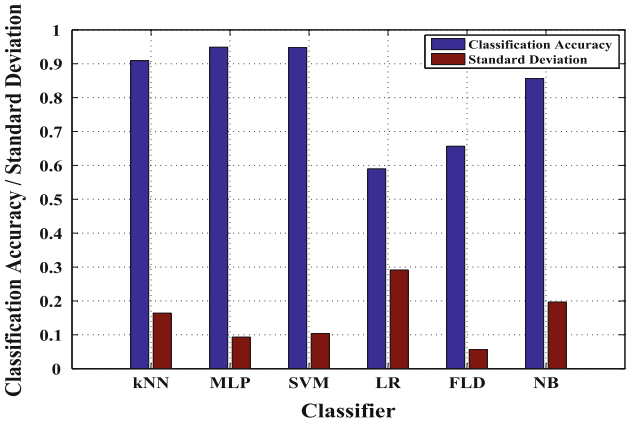

Fig. 7 Classification accuracies and standard deviations from applying *k*NN, single MLP, single SVM, Logistic Regression (LR), Fisher Linear Discrimination (FDL), and Naive Bayesian (NB)

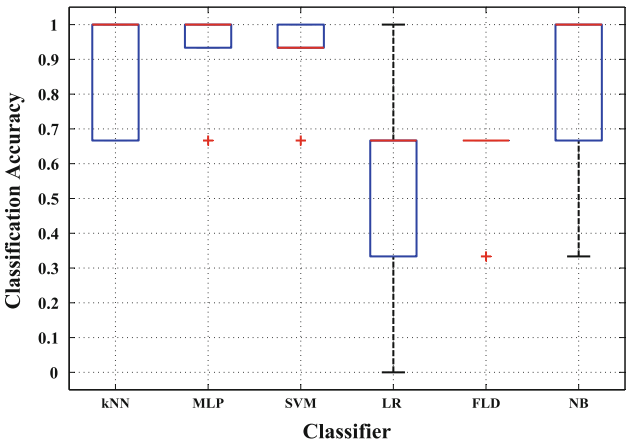

Fig. 8 Boxplot of classification accuracies from applying single MLP, single SVM expert, Random Subspace SVM ensemble (RS-SVM) and Random Subspace MLP ensemble (RS-MLP)

results, and their averaged classification accuracies are quite close to each other. The results here contrast to the generally accepted perception that SVM classifiers outperform neural network classifiers. The most reasonable explanation for the better performance of MLP with respect to our experiment is that MLP, as a memory-based classifier, is more resistant to errors introduced from insufficient data than the margin or distance-based SVM. Given a limited amount of data, Naive Bayesian classifier, Linear Discriminant and Logistic Regression perform worse than SVM and MLP, this is because these classifiers' performances depends on the amount of training data, correlations between features, and the probability distribution of each feature, which may vary with empirical data. This part of the experiment demonstrated a common result, also obtained with respect to other research work, that in general SVM and MLP can achieve better classification performance on biopsy image analysis.

4.2 Evaluation of random subspace ensembles

Table 2 shows the classification accuracies obtained using 7 different ensemble classifiers with different image feature combinations. The classifier ensemble methods compared here are: (1) Bagging [50] with SVM (BagSVM), (2) Bagging with MLP (BagMLP), (3) AdaBoost [51] with SVM (BoostSVM), (4) AdaBoost with MLP (BoostMLP), (5) Random Forest [52] with decision trees (RandF), (6) Random Subspace with MLP (RSMLP) and (7) Random Subspace with SVM (RSSVM). The three different image feature types introduced earlier were considered: Curvelet, GLCM, and LBP, they are represented by the letters C, G, and L in Table 2, respectively. Each image has a 666-dimensional feature vector with all of these three features. Each randomly selected subspace used 80 % of the features for the training phase of the classifiers. For example, a 532-dimensional (666×0.8) feature vector is used for training when three kinds of features are all used (C, G and L in Table 2). In order for comparison, the full (100 %) feature vectors were also used for classifier training, the results of using full feature vectors are listed in the last column of the table. The ensemble size is fixed as 25 for all the classifiers in Table 2.

One can note from Table 2 that the use of ensembles does improve the classification accuracy. RSSVM and RSMLP produced the best performance, both obtain classification accuracies around 95 % regardless of the types of image features used for the training (Curvelet, GLCM orLBP), which is much better than the results obtained by other feature combinations. The results of the Random Subspace ensemble (RSSVM, RSMLP) using 80 % features for training are also better than the results of using the whole feature vector in the training phase, which means the classification task benefits from Random Subspace ensemble.

The results on this image dataset from using other kinds of features are also compared in the experiment, as in [5], the level set method was used to extract image features, and a 42-bin histogram was constructed to represent information of connected components; a 6.6 % classification error rate was obtained.

Two important parameters for Random Subspace ensembles are ensemble size (L) and the cardinality of the feature vectors (M). A "rule of thumb" has been put forward with respect to the fMRI data classification problem [16] in which the authors proposed a feature subset size $M = \frac{n}{2}$ and a consequent ensemble size of $L = \frac{n}{10}$, where n is the dimension of the original feature vector. In order to find the appropriate values for the ensemble size and feature vector cardinality for the current biopsy image classification work, the size of the ensembles was varied from 5 to 145 with a

Table 2 Classification accuracy (%) of 7 Ensemble classifiers on the biopsy image data with different image feature combinations

Ensemble	Features used							
	C	G	L	C&G	C&L	G&L	C&G&L	100 %
BagSVM	87.56	87.21	88.53	89.65	90.06	90.48	92.04	91.67
BagMLP	87.56	87.42	88.84	90.75	90.58	90.67	93.44	93.02
BoostSVM	86.81	86.06	87.54	89.25	89.54	90.70	92.70	92.88
BoostMLP	87.72	87.21	88.44	90.17	90.22	90.44	93.22	93.56
RandF	82.73	82.61	83.25	85.81	84.61	87.03	89.81	92.44
RSMLP	90.43	90.82	91.79	92.58	93.39	93.89	95.05	94.88
RSSVM	90.13	90.09	90.44	92.08	92.51	92.78	94.85	94.12

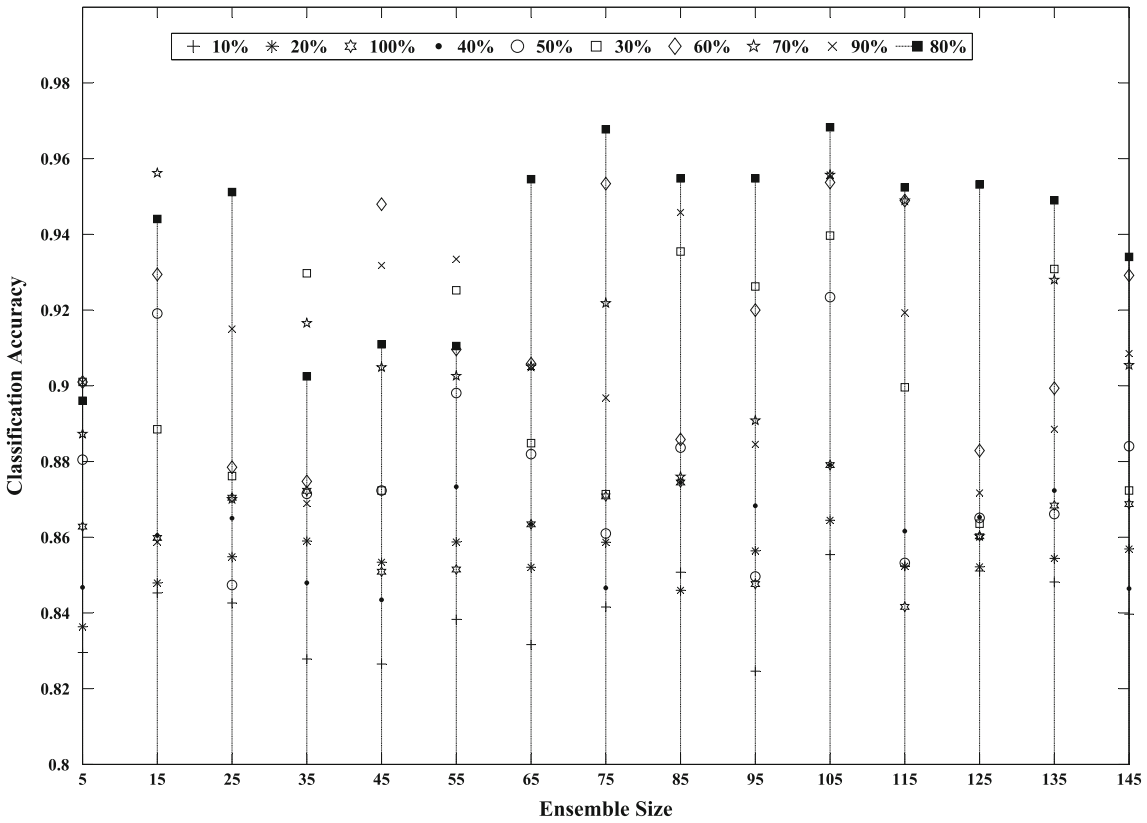

Fig. 9 Classification results of the RSSVM ensemble with different ensemble sizes and different cardinalities of training feature

step size of 10. For each ensemble value size, the cardinality of the feature vectors used for training was changed from 10 % of the original dimension to 100 %, with equally spaced intervals of 10 %. The classification results using RSSVM and RSMLP with different ensemble sizes and different feature vector cardinalities are shown in Figs. 9 and 10, respectively.

The same conclusion as in [24] can be drawn from Figs. 9 and 10, namely that the classification performance does not rely on the increase of the ensemble size. The different cardinalities of the feature vectors produced different performances. The Random Subspace MLP ensemble obtains its best classification accuracy of 96.83 % using $M = \frac{4n}{5}$ and ensemble size $L = 105$. The Random Subspace SVM ensemble also achieved good performance with an accuracy 96.56 % at 80 % feature cardinality; however, different from the MLP ensemble, the SVM ensemble has the same top performance for ensemble sizes 85 to 115. Therefore, the most appropriate feature cardinality of $M = \frac{4n}{5}$ and ensemble size $L = 105$ were identified for both of the Random Subspace MLP ensemble and the SVM ensemble.

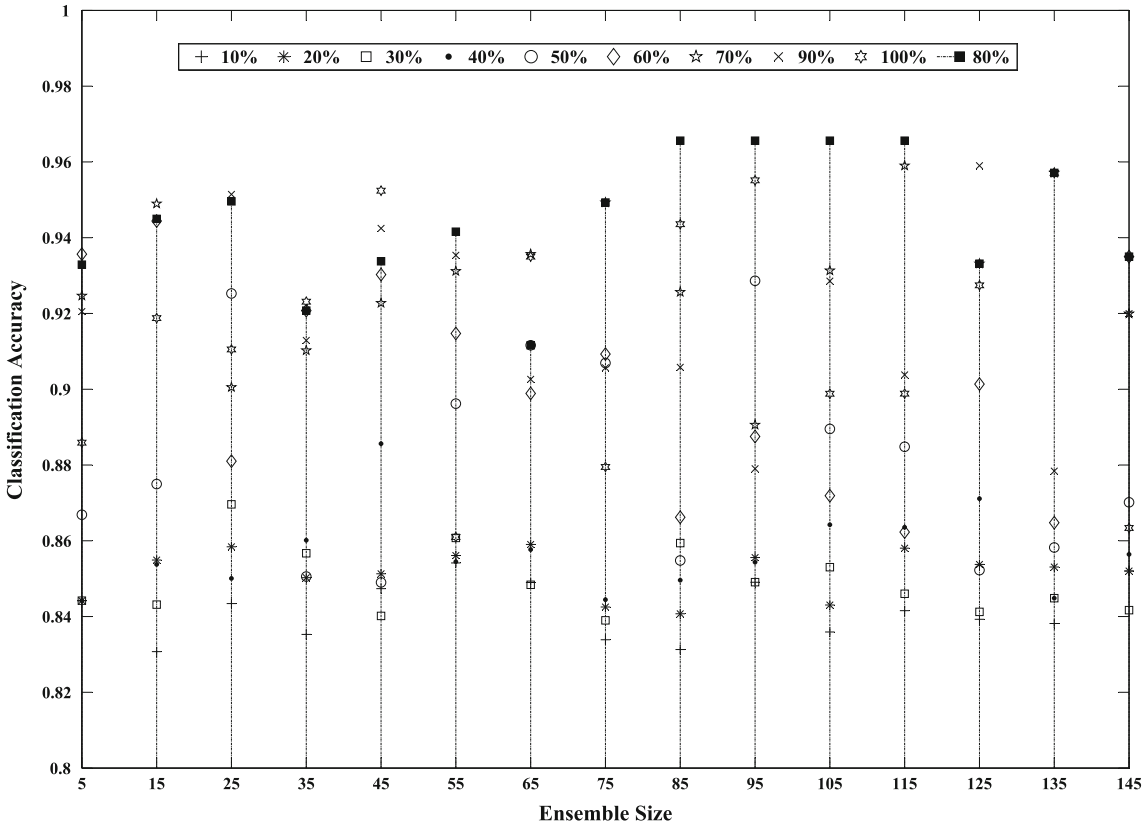

Fig. 10 Classification results of the RSMLP ensemble with different ensemble sizes and different cardinalities of training feature

4.3 Results of the proposed ensemble cascade system

In this experiment, we first use the RSSVM-ensemble and the RSMLP-ensemble to construct different cascade classification systems. Four different two-stage cascade classifiers were built: RSSVM-RSSVM, RSMLP-RSMLP, RSSVM–RSMLP, and RSMLP-RSSVM; where RSSVM-RSSVM indicates that a RSSVM ensemble was employed in both stages 1 and 2, RSSVM–RSMLP indicates that a RSSVM ensemble was used in stage 1 and a RSMLP ensemble in stage 2, and so on.

The parameters for the RSSVM and RSMLP ensembles were as determined in the previous experiment, with ensemble sizes equal to 105 and feature cardinality set to 80 %. A rejection threshold 84 (0.8×105) was set for both ensembles (stages 1 and 2), which means that only when more than 80 % of the classifiers agree on some decision will the decision be adopted, otherwise, the instance will be rejected by the ensemble. This relatively high threshold was used because we wished to insure a high level of reliability with respect to classification decisions. The results of different cascade schemes on the biopsy image dataset are listed in Table 3.

From Table 3, it can be observed that all the two-stage cascade classifiers obtain a better classification performance than the non-cascade ensembles tested in the last experiment, this confirms the effectiveness of the cascade classification system, which benefits from the fact that the samples rejected by the first ensemble still have the chance to be correctly classified by the second ensemble. Among the four different cascade classifiers, the RSSVM–RSMLP cascade classifier obtained the best classification accuracy with a relatively low rejection rate. The reasonable explanation is that use of different base classifiers in the ensembles increase the diversity of the whole cascade system, and compared with SVM, MLP is a more 'localized' classifier which is more suitable to be put in stage 2 to achieve better performance [24].

Table 3 Classification accuracy and reliability of different cascade schemes on the biopsy image data with rejection threshold of both stages equal to 84

Cascades	RR (%)	Re (%)	ReR (%)	ER (%)
RSSVM-RSSVM	97.19	97.63	1.43	2.38
RSMLP-RSMLP	97.39	98.22	1.19	1.78
RSSVM-RSMLP	98.61	98.65	0.53	1.35
RSMLP-RSSVM	97.89	98.40	1.71	1.60

RR recognition rate, *Re* reliability, *ReR* rejection rate, *ER* error rate, see Sect. 3 for details

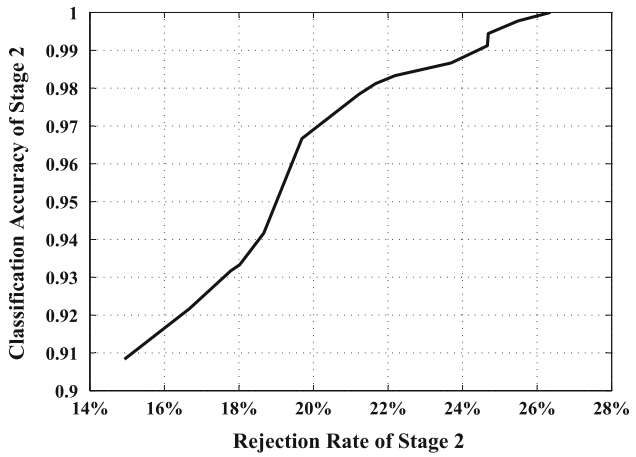

Fig. 11 Averaged stage 2 accuracies with 10 varying stage 2 rejection rates

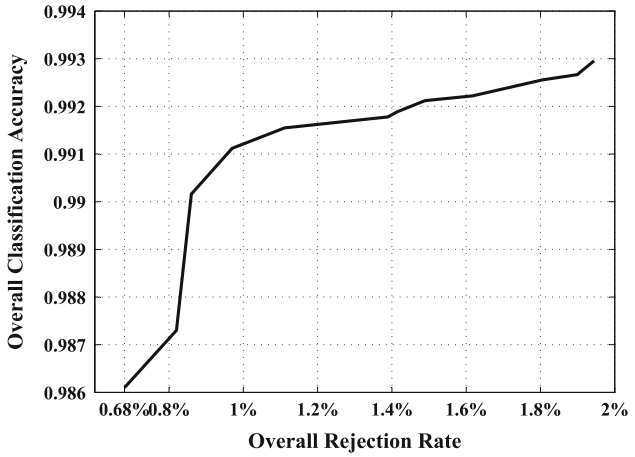

Fig. 12 Averaged overall classification performances from 10 varying overall rejection rates

To have a closer look at how the rejection rate influences the classification accuracy, we adjusted the threshold t_2 for the majority voting of the stage 2 ensemble (t_2-out-of-L, $L = 105$), while fixing the threshold in stage 1 at $t_1 = 84$ (0.80×105), resulting in average rejection rates at stage 2 of between 14.29 and 26.36 % from $t_2 = 85, \ldots, 95$. The corresponding overall rejection rates were then in the range of $0.68, \ldots, 1.94$ %. The plots of stage 2 accuracies and corresponding overall accuracies from the varying rejection rates are displayed in Figs. 11 and 12, respectively. It is not difficult to appreciate that higher accuracy could be expected from higher rejection rate. However, it is worth noting that when the rejection rate of stage 2 is 26.36 %, the classification accuracy of stage 2 is 100 %, as we continued increasing the value of the threshold t_2, the increased rejection rate did not bring any more improvement with respect to the classification performance.

With $t_1 = 84$ and $t_2 = 95$, the classification accuracies and reliabilities from stage 1, stage 2 and the whole system can be seen in Table 4. Compared with the results in Table 3, where the same thresholds $t_1 = t_2 = 84$ was set for both stages, the overall classification accuracy and reliability were improved by increasing the value of t_2, and the corresponding error rate drops. However, this improved performance is obtained at the cost of an augmented rejection rate, which means there will be more images left for human experts to analyze. The trade-off between accuracy and rejection rate could be empirically decided in practice.

The confusion matrix from the overall performance that summarize the detailed situations of rejection rate 1.94 % were displayed in the Table 5. In the confusion matrix representation, the rows and columns indicate the true and predicted classes, respectively. The diagonal entries represent correct classification while the off-diagonal entries represent incorrect ones.

Table 4 Averaged classification performance of the cascade schemes on the biopsy image data with rejection threshold $t_1 = 84$ and $t_2 = 95$

	RR (%)	Re (%)	ReR (%)	ER (%)
Stage 1 (RSSVM)	98.61	99.31	7.73	0.69
Stage 2 (RSMLP)	1	83.64	26.36	0
Cascade	99.25	97.65	1.94	1.25

Table 5 Averaged confusion matrix with overall rejection rate 1.94 % (%)

	Insitu	Normal	Invasive
Insitu	97.97	0.74	1.29
Normal	0	100	0
Invasive	0.22	0	99.78

5 Conclusion and future work

In this paper, a reliable classification scheme based on serial fusion of Random Subspace ensembles has been proposed for the classification of microscopic biopsy images for breast cancer diagnosis. Rather than simply pursuing classification accuracy, we emphasized the importance of a reject option in order to minimize the cost of misclassifications so as to ensure high classification reliability. The proposed two-stage method used a serial approach where the second classifier ensemble is only responsible for the patterns rejected by the first classifier ensemble. The first stage ensemble consists of binary SVMs which were trained in parallel, while the second ensemble comprises MLPs. During classification, the cascade of classifier ensembles received randomly sampled subsets of features following the Random Subspace procedure. For both of the ensembles the rejection option was implemented by relating the consensus degree from majority

voting to a confidence measure and abstaining to classify ambiguous samples if the consensus degree was lower than the threshold.

The effectiveness of the proposed cascade classification scheme was verified on a breast cancer biopsy image dataset. The combined feature representation from LBP texture description, Gray Level Co-occurrence Matrix and Curvelet Transform exploits the complementary strengths of different feature extractors; the combined feature was proved efficient with respect to the biopsy image classification task. The two-stage ensemble cascade classification scheme obtained a high classification accuracy (99.25 %) and simultaneously guaranteed a high classification reliability (97.65 %) with a small rejection rate (1.94 %). Moreover, the cascade architecture provides a mechanism to balance between classification accuracy and rejection rate. By adjusting the rejection threshold in each ensemble, the classification accuracy and reliability of the system can be modulated to a certain degree according to the specification of specific applications. For example, medical diagnosis tasks usually require high accuracy and reliability, therefore the rejection thresholds in each stage will be set to a high level in order to guarantee the correctness of the diagnosis.

Although the proposed system has shown promising results with respect to the biopsy image classification task, there are still some aspects that need to be further investigated. The benchmark images used in this work were cropped from the original biopsy scans and only cover the important areas of the scans. However, often it is difficult to find Regions of Interest (ROIs) that contain the most important tissues in biopsy scans, more efforts therefore needs to be put into detecting ROIs from biopsy images. In this paper, the parameters for the cascade system, such as ensemble size and rejection threshold, were decided empirically; this may not have produced the most satisfactory performance with respect to all application contexts. Therefore, some self-adaptive rules or algorithms for automatically optimizing these parameters would be desirable.

References

1. Breast Cancer Facts & Figures 2009–2010, American Cancer Society (2010)
2. Gaurav, A., Pradeep, P.V., Aggarwal, V., Yip, C.-H., Cheung, P.S.Y.: Spectrum of breast cancer in Asian women. World J. Surg. **31**(5), 1031–1040 (2007)
3. Linos, E., Spanos, D., Rosner, B.A., Linos, K., Hesketh, T., Qu, J.D., Gao, Y.-T., Zheng, Wei, Colditz, Graham A.: Effects of reproductive and demographic changes on breast cancer incidence in china: a modeling analysis. J. Natl. Cancer Inst. **100**(19), 1352–1360 (2008)
4. Arisio, R., Cuccorese, C., Accinelli, G., Mano, M.P., Bordon, R., Fessia, L.: Role of fine-needle aspiration biopsy in breast lesions: analysis of a series of 4,110 cases. Diagn. Cytopathol. **18**(6), 462–467 (1998)
5. Brook, A., El-Yaniv, R., Isler, E., Kimmel, R., Meir, R., Peleg, D.: Breast cancer diagnosis from biopsy images using generic features and SVMs. Tech. Rep. CS-2008-07, Technion-Israel Institute of Technology, Technion City, Haifa 32000, Isreal (2006)
6. Boucheron, L.E.: Object- and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer. Ph.D. Thesis, University of California Santa Barbara, Santa Barbara, CA (2008)
7. Loukas, C.: A survey on histological image analysis-based assessment of three major biological factors influencing radiotherapy: proliferation, hypoxia and vascluature. Comput. Methods Programs Biomed. **74**(3), 183–199 (2004)
8. Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D.M., Goldberg, I.G.: Wnd-charm: multi-purpose image classification using compound image transforms. Pattern Recognit. Lett. **29**(11), 1684–1693 (2008)
9. Tabesh, A., Teverovskiy, M., Pang, H.-Y., Kumar, V.P., Verbel, D., Kotsianti, A., Saidi, O.: Multifeature prostate cancer diagnosis and gleason grading of histological images. IEEE Trans. Med. Imaging **26**(10), 1366–1378 (2007)
10. Qureshi, H., Sertel, O., Rajpoot, N., Wilson, R., Gurcan, M.: Adaptive discriminant wavelet package transform and local binary patterns for meningioma subtype classification. MICCAI **2008**, 196–204 (2008)
11. Gurcan, Metin N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B.: Histopathological image analysis: a review. IEEE Rev. Boimed. Eng. **2**, 147–171 (2009)
12. Yang, P., Yang, Y.H., Zhou, B.B., Zomaya, A.Y.: A review of ensemble methods in bioinformatics. Curr. Bioinforma. **5**(4), 296–308 (2010)
13. Freund, Y.: Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2), 256–285 (1995)
14. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, New York (2004)
15. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. PAMI **20**, 832–844 (1998)
16. Kuncheva, L.I., Rodriguez, J.J., Plumpton, C.O., Linden, D.E., Johnston, S.J.: Random subspace ensembles for FMRI classification. IEEE Trans. Med. Imaging. **29**(2), 531–542 (2010)
17. Bertoni, A., Folgieri, R., Valentini, G.: Biological and Artificial Intelligence Environments. Springer, Berlin (2008)
18. Pudil, P., Novovicova, J., Blaha, S., Kittler, J.: Multistage pattern recognition with reject option. In: Proceeding of the Eleventh IAPR International Conference on Pattern Recognition B, pp. 92–95 (1992)
19. Alpaydin, E., Kaynak, C.: Cascading Classifiers. Kybernetika, vol. 34, pp. 369–374
20. Kaynak, C.: Alpaydin, E.: Multistage cascading of multiple classifiers: one man's noise is another man's data. In: Proceedings of ICML, 2000, pp. 455–462 (2000)
21. Chow, C.K.: On optimum recognition error and reject tradeoff. IEEE Trans. Inf. Theory **IT–16**(1), 41–46 (1970)
22. Pudil, P., Novovicova, J., Blaha, S., Kittler, J.: Multistage pattern recognition with reject option. In: Proceedings of 11th IAPR International Conference on Pattern Recognition, vol. **2**, pp. 92–95 (1992)
23. Fumera, G., Roli, F.: Support vector machines with embedded reject option. In: International Workshop on Pattern Recognition with Support Vector Machines (SVM2002), pp. 68–82. Springer, Niagara Falls, Canada (2002)
24. Giusti, N., Masulli, F., Sperduti, A.: A theoretical and experimental analysis of a two-stage system for classification. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 893–904 (2002)

25. Nadeem, MSA., Zucker, J.-D., Hanczar, B.: Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In: Proceedings of the Third International Workshop on Machine Learning in Systems Biology, Ljubljana, Slovenia, pp. 5–6 (2009)
26. Hanczar, B., Dougherty, E.R.: Classification with reject option in gene expression data. Bioinformatics **24**(17), 1889–1895 (2010)
27. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (July 2002)
28. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Process. (2010) (accepted)
29. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. **3**(6), 610–621 (1973)
30. Boland, M.V.: Quantitative description and automated classification of cellular protein localization patterns in fluorescence microscope images of mammalian cells. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh (1999)
31. Clausi, D.A.: An analysis of co-occurence texture statistics as a function of grey level quantization. Can. J. Remote Sensing **28**(1), 45–62 (2002)
32. Soh, L.-K., Tsatsoulis, C.: Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Trans. Geosci. Remote Sensing **37**(2), 780–795 (1999)
33. Donoho, D., Duncan, M.: Digital Curvelet Transform: Strategy. Implementation and Experiments. Stanford University, Stanford (1999)
34. Starck, J., Candes, E., Donoho, D.: The curvelet transform for image denoising. IEEE Trans. Image Process. **11**, 670–684 (2002)
35. Candes, E., Donoho, D.: Curvelets: multiresolution representation, and scaling laws. In: Aldroubi, A., Laine, A.F., Unser, M.A. (eds.) Wavelet Applications in Signal and ImageProcessing VIII, Proceeding of the SPIE 4119 (2000)
36. Candes, E., Demanet, L., Donoho, D., Ying, L.: Fast discrete curvelet transforms. Multiscale Model. Simul. **5**, 861–899 (2006)
37. Ma, J., Plonka, G.: The curvelet transform: a review of recent applications. IEEE Signal Process. Mag. **27**(2), 118–133 (2010)
38. Meselhy Eltoukhy, M., Faye, I., Belhaouari Samir, B.: Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. Comput. Med. Imaging Graph. **34**, 269–276 (2010)
39. Zhang, B., Pham, T.D.: Phenotype recognition with combined features and random subspace classifier ensemble. BMC Bioinforma. **12**, 128 (2010)
40. Zhang, P., Bui, T.D., Suen, C.Y.: A novel cascade ensemble classifier system with a high recognition performance on handwritten digits. Pattern Recognit. **40**, 3415–3429 (2007)
41. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)
42. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
43. Kuncheva, L.I., Skurichina, M., Duin, R.P.: An experimental study on diversity for bagging and boosting with linear classifiers. Inf. Fusion **3**(4), 245–258 (2002)
44. Tax, D.M.J., Duin, R.P.W.: Growing a multi-class classifier with a reject option. Pattern Recognit. Lett. **29**(10), 1565–1570 (2008)
45. Tax, D.M.J., Duin, R.P.W.: Support vector domain description. Pattern Recognit. Lett. **20**(11–13), 1191–1199 (1999)
46. Lam, L., Suen, C.Y.: Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Trans. Syst. Man Cybern. Part A Syst. Human **27**, 553–568 (1997)
47. Giusti, N., Masuli, F., Sperduti, A.: Theoretical and experimental analysis of a two-stage system for classification. IEEE Trans. PAMI **24**(7), 893–904 (2002)
48. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
49. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
50. Breiman, L.: Bagging predictors. Mach. Learn. **26**(2), 123–140 (1996)
51. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceeding of 13th International Conference on Machine Learning, San Francisco, CA, USA. pp. 148–156 (1996)
52. Breiman, L.: Random For. Machine Learning **45**, 5–32 (2001)