Cascade of Classifier Ensembles for Reliable Medical Image Classification

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by
Yungang Zhang

March 2014

Declaration

1. Candidate's declarations:

I, Yungang Zhang, hereby certify that this thesis, which is approximately 43,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

Date: 4 March 2014. **Signature:** Yungang Zhang .

2. Supervisor's declarations:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor in Philosophy in the University of Liverpool and that the candidate is qualified to submit this thesis in application for that degree.

Date: 4 March 2014. **Signature:** Wenjin Lu .

Acknowledgement

I would like to express my deepest appreciation and sincere gratitude to my supervisors, Dr. Wenjin Lu, Dr. Bailing Zhang and Professor Frans Coenen for their invaluable advice and great support in my study. It is them who encourage me to get through the tough time during the research. I not only learned specialized knowledge from them, but also the attitude and spirit of exploring scientific problems. This work would not have been possible without their excellent guidance and persistent encouragement.

I appreciate the Department of Computer Science and Software Engineering, Xi-anJiaoTong Liverpool University (XJTLU) and the Department of Computer Science, University of Liverpool for kindly awarding me the research scholarship and providing excellent facilities for my research.

I would like to thank Dr. Ting Wang, Mr. Jieming Ma and other research students and staff in XJTLU for their valuable suggestions and discussions. In addition, I am grateful to Mr. Abdulrahman Albarrak from Department of Computer Science, University of Liverpool, and Dr. Yalin Zheng from Royal Hospital of University of Liverpool for their unselfish support of the 3D OCT retina image data.

Finally, I would like to dedicate this thesis to my family and my wife, Lihui Wei, for their unlimited support and love.

Publications

1. Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu. Breast Cancer Diagnosis from Biopsy Images with Highly Reliable Random Subspace Classifiers Ensemble. *Machine Vision and Applications* 24(7): 1405-1420, 2013.

2. Jimin Xiao, Tammam Tillo, Chunyu Lin, Yungang Zhang and Yao Zhao. A Real-Time Error Resilient Video Streaming Scheme Exploiting the Late- and Early-Arrival Packets. *IEEE Transactions on Broadcasting* 59(3): 432-444, 2013.

3. Yungang Zhang, Bailing Zhang, Wenjin Lu, Breast Cancer Histological Image Classification with Multiple Features and Random Subspace Classifier Ensemble. T.D. Pham, L.C. Jain (eds): *Innovations in Knowledge-based Systems in Biomedicine*, Springer-Verlag, SCI 450, pp. 27-42, 2013. (book chapter).

4. Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu, One-Class Kernel Subspace Classifier Ensemble for Biopsy Image Classification, *Eurasip Journal on Advances in Signal Processing*, in revision.

5. Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu, Cascading One-Class Kernel Subspace Ensembles for Reliable Biopsy Image Classification. Accepted by *Journal of Medical Imaging and Health Informatics*.

6. Bailing Zhang, Yungang Zhang, Wenjin Lu, and Guoxia Han. Phenotype Recognition by Curvelet Transform and Random Subspace Ensemble. *Journal of Applied Mathematics & Bioinformatics*, vol.1, no.1, pp. 79-103, 2011.

7. Yungang Zhang, Tianwei Xu and Wei Gao, Image Retrieval Based on GA Integrated Color Vector Quantization and Curvelet Transform. In *Proceedings of 2012 International Conference of Swarm Intelligence (ICSI2012)*, pp. 406-413. Shenzhen, China, 2012.

8. Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu, Highly Reliable Breast Cancer Diagnosis with Cascaded Ensemble Classifiers. In *Proceedings of 2012 International Joint Conference on Neural Networks (IJCNN 2012)*. Brisbane, Australia, June 2012.

9. Yungang Zhang, Wei Gao, and Jun Liu. Integrating Color Vector Quantization and Curvelet Transform for Image Retrieval. *International Journal of Design, Analysis and Tools for Circuits and Systems*, vol. 2, no. 2, pp.99-106, 2011.

10. Yungang Zhang, Bailing Zhang, Wenjin Lu, Breast Cancer Classification From Histological Images with Multiple Features and Random Subspace Classifier Ensemble. *In Proceedings of CMLS 2011, AIP Conf. Proc.* Volume 1371, pp. 19-28, Toyama, Japan, June 2011.

11. Yungang Zhang, Lijin Gao, Wei Gao and Jun Liu, Combining Color Quantization with Curvelet Transform for Image Retrieval. *In Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI2010)*, Vol.1, pp. 474-479, Sanya, China, 2010.

12. Yungang Zhang, Bailing Zhang, and Wenjin Lu. Image Denoising and Enhancement Based on Adaptive Wavelet Thresholding and Mathematical Morphology. *In Proceedings of 2010 IEEE International Congress on Image and Signal Processing (CISP2010)*, pp. 973-976. Yantai, China, October 2010.

Abstract

Medical image analysis and recognition is one of the most important tools in modern medicine. Different types of imaging technologies such as X-ray, ultrasonography, biopsy, computed tomography and optical coherence tomography have been widely used in clinical diagnosis for various kinds of diseases. However, in clinical applications, it is usually time consuming to examine an image manually. Moreover, there is always a subjective element related to the pathological examination of an image. This produces the potential risk of a doctor to make a wrong decision. Therefore, an automated technique will provide valuable assistance for physicians. By utilizing techniques from machine learning and image analysis, this thesis aims to construct reliable diagnostic models for medical image data so as to reduce the problems faced by medical experts in image examination. Through supervised learning of the image data, the diagnostic model can be constructed automatically.

The process of image examination by human experts is very difficult to simulate, as the knowledge of medical experts is often fuzzy and not easy to be quantified. Therefore, the problem of automatic diagnosis based on images is usually converted to the problem of image classification. For the image classification tasks, using a single classifier is often hard to capture all aspects of image data distributions. Therefore, in this thesis, a classifier ensemble based on random subspace method is proposed to classify microscopic images. The multi-layer perceptrons are used as the base classifiers in the ensemble. Three types of feature extraction methods are selected for microscopic image description. The proposed method was evaluated on two microscopic image sets and showed promising results compared with the state-of-art results.

In order to address the classification reliability in biomedical image classification problems, a novel cascade classification system is designed. Two random subspace based classifier ensembles are serially connected in the proposed system. In the first stage of the cascade system, an ensemble of support vector machines are used as the base classifiers. The second stage consists of a neural network classifier ensemble. Using the reject option, the images whose classification results cannot achieve the predefined rejection threshold at the current stage will be passed to the next stage for further consideration. The proposed cascade system was evaluated on a breast cancer biopsy image set and two UCI machine learning datasets, the experimental results showed that

the proposed method can achieve high classification reliability and accuracy with small rejection rate.

Many computer aided diagnosis systems face the problem of imbalance data. The datasets used for diagnosis are often imbalanced as the number of normal cases is usually larger than the number of the disease cases. Classifiers that generalize over the data are not the most appropriate choice in such an imbalanced situation. To tackle this problem, a novel one-class classifier ensemble is proposed. The Kernel Principle Components are selected as the base classifiers in the ensemble; the base classifiers are trained by different types of image features respectively and then combined using a product combining rule. The proposed one-class classifier ensemble is also embedded into the cascade scheme to improve classification reliability and accuracy. The proposed method was evaluated on two medical image sets. Favorable results were obtained comparing with the state-of-art results.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Computer-aided diagnosis (CAD) aims to assist medical physicians for making diagnostic decisions with computers. As an interdisciplinary research area, CAD covers technologies in signal processing, pattern recognition, computer vision and machine learning. Medical imaging is one of the most important tools in modern medicine, different types of imaging technologies such as X-ray imaging, ultrasonography, biopsy imaging, computed tomography, and optical coherence tomography have been widely used in clinical diagnosis for various kinds of diseases. However, in clinical applications, it is usually time consuming to examine an image manually. Moreover, there is always a subjective element related to the pathological examination of an image, this produces the potential risk for a doctor to make a wrong decision. Therefore, an automated technique will provide valuable assistance for physicians. By utilizing techniques from machine learning and image analysis, this research aims to construct reliable diagnostic models for medical image data to relieve the problems faced by medical experts in image examination. Through supervised learning of the image data, the diagnostic model can be constructed automatically and then applied in disease diagnosis.

The process of image examination by human experts is very difficult to simulate, as the knowledge of medical experts is often fuzzy and not easy to be quantified. Therefore, the problem of automatic diagnosis based on images is usually converted to the problem of image classification. Feature extraction is the process of creating a representation for the original image data. By extracting the image features which are suitable to indicate the symptoms of diseases, the quantization of medical knowledge can be realized. The different image feature degrees related to different disease situations can be used to train a classifier, then the trained classifier will be able to categorize new image cases. In this research, different image feature descriptors are investigated and combined to produce effective and efficient description for typical types of medical images.

A great number of machine learning methods have been proposed to design accurate classification systems for various medical images. Among them, ensemble learning has

attracted much attention due to good performance from many applications in medicine and biology. Ensemble learning is concerned with mechanisms to combine the results of a number of classifiers. In the case of ensemble classification, ensemble learning is concerned with the integration of the results of a number of classifiers (often called 'base classifiers') to develop a strong classifier with good generalization performance. In this research, ensemble learning strategies are investigated in medical image classification schemes to improve the classification performance.

In previous studies of medical image classifications [161, 68], 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.

Cascading 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. Using 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 and singular points. Many cascading multi-stage classifier architectures have been proposed and plenty of promising results have been achieved in medical and biological classification applications [185]. This motivates the development of new cascade classification schemes to address both classification accuracy and reliability. In this thesis, a two-stage cascade classification model is constructed; each stage in the cascade includes a classifier ensemble. Such a classification model takes advantages of both ensemble learning and cascading so that it can improve classification accuracy and reliability simultaneously.

One challenge in many automatic medical diagnosis applications is that the datasets used for diagnosis are often imbalanced. As the number of normal cases is usually much larger than the number of disease cases, classifiers that generalize well over the balanced data may not be the most appropriate choice in such an unbalanced situation. For example, decision trees tend to over-generalize the class with the most examples; Naive Bayes requires enough data for the estimation of the class-conditional probabilities [119]. One-Class Classifiers (OCC) [192] are more appropriate for such a task. One-class classification is also often called outlier (or novelty) detection as the learning algorithms

are used to differentiate between data that appears normal and abnormal with respect to the distribution of the training data. One-class classification is appropriate with respect to medical diagnosis, i.e., disease versus no-disease problems, where the training data tends to be imbalanced. This motivates us further to develop new types of cascade classification algorithms, which exploit one-class classifiers to tackle the imbalanced data problem, together with the ensemble and cascade learning strategies. Such a cascade classification scheme is expected to improve the classification performance in many medical image classification applications.

1.2 Objectives

The major objective of this research is to develop and evaluate new classification schemes to improve classification accuracy and reliability of many medical image diagnosis applications, such as breast cancer biopsy image classification, 3D OCT retina image classification and fluorescence microscope image classification.

The following aspects of medical image classification problem are investigated and discussed in the thesis:

- Random subspace classifier ensemble for biomedical image classification.
- The cascade classification scheme for reliable medical image classification.
- One-class classifier ensemble to tackle with the imbalanced data distribution in medical image diagnosis.
- Effective image feature description methods for microscopic images.

These novel techniques were implemented and evaluated using the benchmark biomedical image datasets described in the following section (Section 1.3).

1.3 Datasets and Evaluation Methods Used in the Thesis

1.3.1 Datasets

- Three benchmark fluorescence microscopy image datasets in [113] were used in our study, which are RNAi, CHO and 2D-Hela.

  The RNAi dataset is a set of fluorescence microscopy images of fly cells (D. melanogaster) subjected to a set of gene-knockdowns using RNAi. The cells are stained with DAPI to visualize their nuclei. Each class contains 20 1024×1024 images of the phenotypes resulting from knockdown of a particular gene. Ten genes were selected, and their gene IDs are used as class names. The genes are CG1258, CG3733, CG3938, CG7922, CG8114, CG8222, CG 9484, CG10873, CG12284, CG17161.

2D HeLa dataset, a collection of HeLa cell immunofluorescence images containing 10 distinct subcellular location patterns. The subcellular location patterns in these collections include endoplasmic reticulum (ER), the Golgi complex, lysosomes, mitochondria, nucleoli, actin microfilaments, endosomes, microtubules, and nuclear DNA. The 2D HeLa image dataset is composed of 862 single-cell images, each with size 382×512.

CHO is a dataset of fluorescence microscope images of CHO (Chinese Hamster Ovary) cells. The images were taken using 5 different labels. The labels are: anti-giantin, Hoechst 33258 (DNA), anti-lamp2, anti-nop4, and anti-tubulin. The CHO dataset is composed of 340 images, each with size 512×382.

- A breast cancer benchmark biopsy image dataset from the Israel Institute of Technology [1]. 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 5μ 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.

- A 3D OCT retinal image set was collected at the Royal Hospital of University of Liverpool, the image set contains 140 volumetric OCT images, in which 68 images from normal eyes and the remainders are from eyes have Age-related Macular Degeneration (AMD).

- Two datasets from UCI machine learning repository (archive.ics.uci.edu/ml/): Breast cancer Wisconsin and Heart disease.

  The Wisconsin breast cancer image sets were obtained from digitized images of fine needle aspirate (FNA) of breast masses. They describe characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. The 569 images in the dataset are categorized into two classes: benign and malignant.

  The Heart disease dataset contains of 270 patterns, each pattern has 13 attributes. The dataset consists of two categories: normal and disease.

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

1.3.2 Evaluation Methods

The performance metrics used in the thesis are listed as follows:

- Classification Accuracy = number of correctly recognized images / number of testing images.
- Recognition rate (RR) = number of correctly recognized images / (number of testing images - number of rejected images).
- Rejection rate (RejR) = number of rejected images /number of testing images.
- Reliability (RE) = (number of correctly recognized images + number of rejected images)/ number of testing images.
- Error rate (ER): = 100% - reliability.
- ROC: Receiver Operating Characteristic graph.
- AUC: Area under an ROC curve.

1.4 Major Contribution of the Thesis

In this thesis, the random subspace method [78] for classifier ensemble is used for combining different classifiers trained by multiple image features, and a new cascade classification scheme with reject option is developed to improve the classification accuracy and reliability for medical image classification. In order to address the problem of imbalanced data in many medical image diagnosis applications, a new ensemble of one-class classifiers is developed, where the reject option is also included to construct a cascade classifier. The proposed methods were evaluated on several real medical imaging applications and benchmark medical image datasets.

The major contributions of this thesis are summarized as follows:

- A novel automatic microscope image classification scheme based on multiple features and random subspace classifier ensemble. The image features are extracted using the Curvelet Transform, statistics of Gray Level Co-occurrence Matrix (GLCM) and the Completed Local Binary Patterns (CLBP), respectively. The three different features are combined together and used for the random subspace ensemble generation, with a set of neural network classifiers aggregated for producing the final decision. Experimental results on the phenotype recognition from three benchmark fluorescence microscopy image sets (RNAi, CHO and 2D Hela) and a benchmark breast cancer biopsy image set show the effectiveness of the proposed approach. The ensemble model produces better performance compared to any of individual neural networks (Multi-Layer Perceptron, MLP).

This part of our work is described in Chapter 3 of the thesis. The work can also be seen in our published papers [227, 228, 218].

- A new cascade classification scheme of Random Subspace ensembles with reject options is proposed. 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 (SVM) 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 (MLP) ensemble, that focuses on the rejected samples from the first ensemble. For both of the ensembles, the rejection 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 a predefined threshold. The proposed cascade system was evaluated on a benchmark microscopic biopsy image dataset and two UCI machine learning benchmark datasets.

  This part of the work is described in Chapter 4 of the thesis. The work can also be seen in the published papers [223, 224].

- A new cascade classifier ensemble is proposed, with the prospective of one-class classification to address the imbalanced data distribution in medical applications.

  The first ensemble consists of a set of Kernel Principle Component Analysis (KPCA) one-class classifiers trained for each image class with different image features. The second ensemble consists of a Random Subspace Support Vector Machine (SVM) ensemble, that focuses on the rejected samples from the first ensemble. For both of the ensembles, the reject option is implemented so that an ensemble abstains from classifying ambiguous samples if the consensus degree is lower than a threshold. The proposed system was evaluated on a benchmark biopsy image dataset and a 3D OCT retinal image dataset.

  This part of the work is described in Chapter 5 of the thesis. The work can also be seen in the published papers [226, 225].

1.5 Organization of the Thesis

The thesis is organized as follows:

- In Chapter 2, a review of classifier ensemble methods, classification with reject option and one-class classification is presented. Section 2.2 introduces the theory of combining multiple classifiers and some popular classifier ensemble methods. A review of classification with reject option is given in Section 2.3, where the

multi-stage (cascade) classifiers are also introduced. In Section 2.4, an overview of the one-class classification is given.

- Chapter 3 presents the applications of multiple image features and Random Subspace ensemble of neural networks on microscopic images. The proposed multiple features and classifier ensemble was evaluated on a benchmark biopsy image dataset and microscopic fluorescence images.

- In Chapter 4, a cascade system consisting of two Random Subspace ensembles is introduced. The first stage of the cascade is an ensemble of support vector machines, the second stage contains a neural networks ensemble. Both of the ensembles are constructed by random subspace method. The reject option is employed in the ensembles to improve the classification reliability. The proposed cascade classifier was evaluated on the biopsy image dataset and a real 3D OCT retinal image dataset.

- Chapter 5 describes a cascade classifier, which is built up on the One-Class classification theory to address the imbalanced problem in many medical applications. The first stage of the cascade is an ensemble of one-class classifiers and the second stage is an "one-versus-all" SVM ensemble. The proposed system was also evaluated on the biopsy image dataset and the 3D OCT retinal image dataset.

- Conclusions and future work are summarized in Chapter 6.

Chapter 2

Literature Review

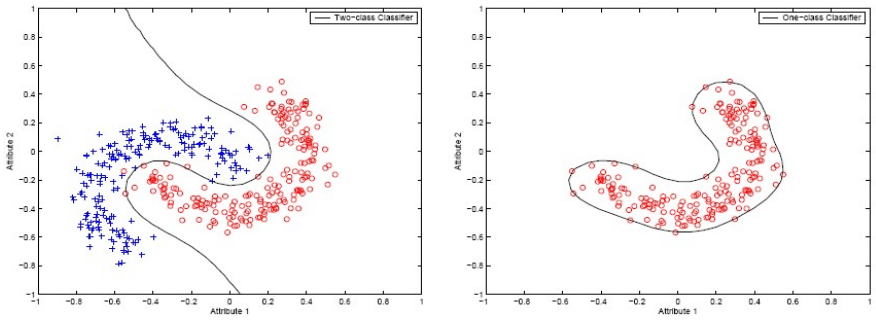
2.1 Introduction

In supervised learning, classification tasks are usually executed by classification models (classifiers), which are constructed from the preclassified instances (samples classified by humans in advance). The preclassfied instances are usually called as *training set*. The goal of the classification model construction is to obtain classifiers from the pre-labeled training sets, then the trained classifiers are able to label the unknown instances.

A number of supervised learning methods have been introduced in the last decades, for example, SVMs [200], neural networks [33], logistic regression [135], naive Bayes [163], random forests [15] and decision trees [135]. The pursuit of higher accuracy has been the main motivation in classifier research. In many real classification tasks, the use of a single classifier often fails to capture all aspects of the data. Therefore, a combination of classifiers (an ensemble) is often considered to be an appropriate mechanism to address this shortcoming. Ensemble learning generates a set of base classifiers using different distributions of training data and then aggregates their outputs to classify new samples [89]. These ensemble learning methods enable users to achieve more accurate predictions with higher generalization abilities than the predictions generated by individual models or experts on average [127].

In recent years, there is a growing demand from many real classification applications that classifiers should have a higher reliability on the classification results. For example, in medical diagnosis applications, making a wrong diagnosis can be very dangerous. Such applications need the classification systems to keep their classification error as low as possible. Accordingly, the classifiers should have the ability to make no judgement on the ambiguous instances. One way to endow a classifier with such an ability is to implement reject option [30].

Classification with a rejection option has been a topic of interest in pattern recognition. Multi-stage classifiers are serial ensembles where individual classifiers have a reject option [151]. Cascading [50] is a scheme to support multi-stage classification. At the first stage of a cascading system, the system constructs a simple rule using a



(a) Two-class classification (b) One-class classification

Figure 2.1: Decision boundaries of (a) two class classifier and (b) one-class classifier.

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 up 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 [95]. The concept of rejection gives the classifiers the ability to postpone the pattern classification than to take the risk of making an error.

One-class classification is also known as novelty detection and outlier detection [189]. Compared with the conventional two-class classification classifiers like SVM, one-class classifiers assume that only the information of one of the classes (the *target class*) is available, and there is no information about other classes (the *outlier class*). In Fig. 2.1 (a), a two class classifier is trained by the data from both two classes, the aim of the classifier is to obtain a classification boundary discriminating the two classes. While in one-class classification scenario, the classifier is trained by the data only from one class (the target class), the goal of the one-class classifiers is to estimate a decision boundary of the target class and exclude the data of the outlier classes as much as possible (Fig. 2.1 (b)).

Like many automatic medical diagnosis applications, the datasets used for diagnosis is often imbalanced as the number of normal cases is usually larger than the number of the disease cases. Moreover, to label the training samples by human experts are costly. Classifiers that generalize well over balanced data are not the most appropriate choice in such an unbalanced situation. One-Class Classifiers (OCC) are more appropriate for such a task. One-class learning algorithms can differentiate between data that appears normal and abnormal. It is thus significant to investigate one-class classification in medical diagnosis, disease versus no-disease problems, where the training data tends to be imbalanced and limited.

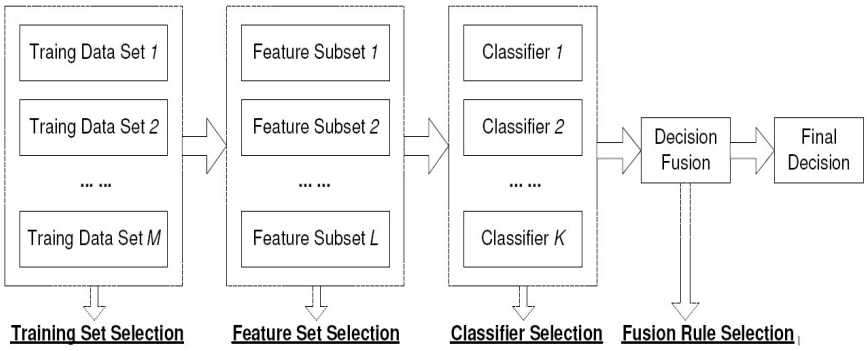


Figure 2.2: Classifier fusion to design an ensemble system

2.2 Ensemble Learning

The idea of ensemble learning was first introduced in the late of 1970's. Two linear regression models were combined to fit the original data and the residuals respectively [201]. The concept of ensemble learning was greatly improved in 1990's, mainly due to the foundation work on boosting [51] and Adaboost algorithm [211], which shows that a strong classifier can be generated by the combination of several weak classifiers. Many researchers have verified the advantages of ensemble learning. Nowadays ensemble learning has been widely used in many pattern recognition applications. Ensemble of classifiers is the focus of this thesis.

2.2.1 Framework of Multiple Classifier Ensemble

Multiple classifier ensemble is also known as mixture of experts, classifier fusion and combination of multiple classifiers, etc [105]. The classifier ensembles aim to combine a set of classifiers to produce better classification performance than each individual classifier can provide. According to Woods et al. [207], a multiple classifier system can be categorized into one of the two categories: classifier fusion or classifier selection. In classifier fusion, the outputs of the individual classifiers are aggregated to make the final decision, the individual classifiers are trained in parallel (Fig 2.2). In classifier selection, only the output of the classifier with the best performance in the ensemble will be selected as the final decision.

Ensemble strategies can be categorized as the dependent framework and independent framework [166]. In a dependent framework, the output of a classifier is used in construction of the next classifier, therefore it is possible to take advantage of knowledge obtained in the previous iterations to guide the learning in the next iterations. Such a framework is called model guided instance selection [177] (Fig. 2.3). In the independent framework, each classifier in the ensemble is built up independently and their results are then combined with some fusion rules.

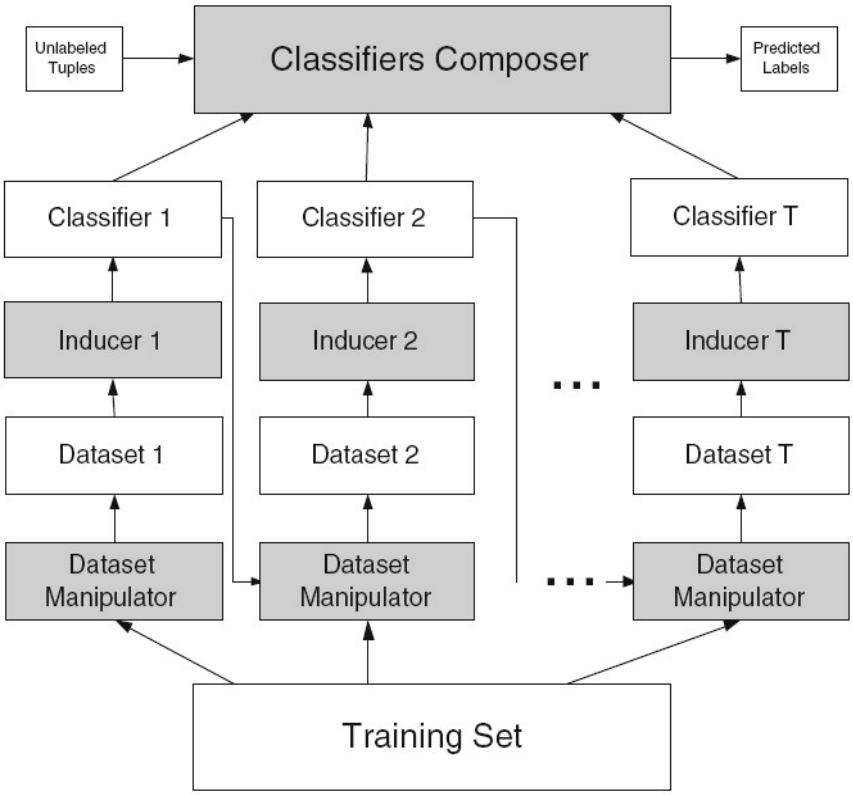


Figure 2.3: Model guided instance selection diagram [166].

2.2.2 Ensemble methods

Boosting

Boosting [51] is the most well known dependent ensemble method based on the resampling technique. Resampling is a widely used technique for generating classifier ensemble. In resampling based ensemble methods, such as boosting and bagging [111], subsets of data are generated for training classifiers, and a learning algorithm is used to obtain multiple predictions on these different training sets. The resampling based methods are effective with unstable classifiers, which are classifiers sensitive to small changes in the training data, such as neural networks and decision trees [53].

Boosting aims to improve the performance of individual classifiers (base classifiers) by repeatedly running the classifiers on various distributed training data. The outputs of individual classifiers are then combined to produce a strong classifier, which is expected to have a better performance than any of the base classifiers.

Freund and Schapire introduced the AdaBoost (adaptive boosting) in 1996 [211]. Compared with the traditional boosting algorithm, AdaBoost tries to improve the final performance by focusing on the patterns that are hard to discriminate. Initially, all the patterns in the training set will be assigned a same weight. Then in each iteration

of the algorithm, the weights of the misclassified patterns will be increased, on the contrary, the weights of the correctly recognized patterns will be decreased. Thus, the base classifiers will give more focus on the hard patterns. At the end of the iterations, each classifier in the ensemble will be assigned a weight, which indicates the overall accuracy of the classifier. The more accurate classifiers will obtain higher weights. The final assigned weights will be used in the classification of new patterns.

The AdaBoost algorithm is first designed to tackle the binary classification problems. Freund and Schapire also proposed two variants of the AdaBoost to address the multiclass classifications. They are named as AdaBoost.M1 and AdaBoost.M2. In AdaBoost.M1, the multiclass classification is achieved by simply aggregating all outputs of the base classifiers. AdaBoost.M2 uses a *label weighting function* to a probability distribution to each training pattern. Thus, the base classifiers will not only obtain a weight distribution of the classifier but a label weight to describe the quality of the hypothesis. The AdaBoost.M2 requires the base classifiers to minimize the *pseudo loss* σ_t, which is a function of the classifier weights and the label weights. A different version of AdaBoost, Real AdaBoost [55], was proposed by Friedman et al. in 2000. By using an additive logistic regression model in a forward stagewise manner, the output class probability is produced from base classifiers.

A distributed version of AdaBoost, P-AdaBoost is developed by Merler et al. in 2007 [133]. Compared with AdaBoost, P-AdaBoost can work on a network of computing nodes. Zhang and Zhang proposed a new boosting-by-resampling version of AdaBoost, which is called the Local Boosting [37]. In the Local Boosting algorithm, for each pattern, a local error is calculated to determine the probability that the pattern should be selected in the next iteration or not. This is different from the AdaBoost, where a global error is calculated at the end of each iteration. By locally investigating each pattern in the training set, the Lobal Boosting is able to filter the noisy patterns, thus acquiring better performance than AdaBoost. Leistner et al. proposed a novel boosting algorithm, On-line GradientBoost [19], which outperformed On-line AdaBoost on standard machine learning problems and common computer vision applications. Bühlmann and Hothorn proposed Twin Boosting [148], which involves a first round of classical boosting followed by a second round of boosting which is forced to resemble the one from the first round. The method has much better feature selection behavior than boosting, particularly with respect to reducing the number of false positives (falsely selected features).

AdaBoost and its variants have achieved great successes in many applications for two reasons:

1. By combining an ensemble of classifiers, the final performance can be improved.

2. The variance of the combined classifier is much lower than the variances of the base classifiers.

However, AdaBoost may still fail to improve the performance of the base classifiers, due to overfitting, which could be induced by a large number of iterations of the algorithm.

Bagging

Bagging is an abbreviation of bootstrap aggregating [111]. The bagging algorithm obtains the final classification result by aggregating the outputs of base classifiers. Each base classifier is trained by a sample in the training set with replacement scheme. The replacement scheme replaces the training sample with a new one in each iteration of training (Algorithm 1). Using the voting strategy, the most often predicted label will be assigned to a pattern. Bagging can usually provide better performance than the individual base classifier, especially when the base classifiers are unstable ones, because Bagging can eliminate the instability of base classifiers.

Algorithm 1 Bagging algorithm

Input:

I: a base classifier
T: the number of iterations
S: the training set
μ: the subsample size

Output:

$\{M_t\}$: the ensemble; $t = 1, ..., T$
$t \leftarrow 1$
Repeat
$s_t \leftarrow$ Sample μ instances from S with replacement
Build classifier M_t using I on s_t
$t++$
until $t > T$

Different from Boosting, Bagging is an independent ensemble method, the base classifiers are trained in parallel. While instances in boosting are selected based on their assigned weights, instances in bagging are chosen with equal probability. In [40], AdaBoost and Bagging were compared in different scenarios, the authors pointed out that, in general Bagging has better performance than AdaBoost, however, in a low noise situation, AdaBoost outperforms Bagging. Skurichina and Duin [126] discovered that Bagging is more appropriate for small training sample sizes, while boosting is better for large training sample sizes.

The trimmed bagging is proposed in [36], which aims to exclude the bootstrapped classification rules that yield the highest error rates, as estimated by the out-of-bag error rate, and to aggregate over the remaining ones. On the basis of numerical experiments, the authors concluded that trimmed bagging performs comparably to standard bagging when applied to unstable classifiers as decision trees, but yields better results when applied to more stable base classifiers, like support vector machines. In [60], the

authors applied an analytical framework for the analysis of linearly combined classifiers to ensembles generated by Bagging. The novel result of the paper is that the authors related the ensemble size with the bagging misclassification probability, thus giving a ground guideline for choosing bagging ensemble size. A new heterogeneous Bagging models [34] were proposed by Coelho and Nascimento in 2008. The model aims at further increasing the diversity levels of the ensemble models produced by Bagging. The authors presented an evolutionary approach for optimally designing Bagging models composed of heterogeneous components. Their experiment results shown that the evolutionary heterogeneous Bagging are matched against standard Bagging with homogeneous components. In a more recent research [209], Bagging and Boosting are used for constructing ensembles in machine translation systems. A Negative Boostrap model was proposed by Li et al. [117] to tackle the visual categorization problem. Given a visual concept and a few positive examples, the Negative Boostrap algorithm iteratively finds relevant negatives. In each iteration, a small proportion of many user-tagged images are used for training, yielding an ensemble of meta classifiers. Compared with the state-of-the-art, the authors obtained better performance.

Random Forest

Algorithm 2 gives the pseudo-code of random forest [15]. A random forest is constructed from a number of decision trees. Each decision tree is trained by a randomly chosen proportion of attributes of the training instances. The classification of a new instance is given by majority voting.

Algorithm 2 The random forest algorithm

Input:

IDT: a decision tree
T: the number of iterations
S: the training set
μ: the subsample size
N: Number of attributes used in each node

Output:

$\{M_t\}$: the forest; $t = 1, ..., T$
$t \leftarrow 1$
Repeat
$s_t \leftarrow$ Sample μ instances from S with replacement
Build classifier M_t using IDT on $s_t(N)$
$t++$
until $t > T$

Random forest was first designed for decision trees, but it can also be used for other classifiers. The two advantages of random forest make it a popular ensemble method. The first is the efficiency of the algorithm; the second one is the good scalability as it can handle large attributes data.

Random forest have been widely used in various fields, for example, segmentation of video objects [26], computed tomography data analysis [161], protein disorder detection [73], spatial context modeling for visual discrimination [142] and human action detection [214].

Diversity Based Methods

Many researchers in ensemble learning have a consensus that diversity is an important factor to obtain a successful ensemble [108]. Base classifiers with wide diversity can lead to uncorrelated classification results, which can improve the performance of an ensemble. The diversity generation methods can be mainly divided into two classes: diversity generation from base classifier manipulation and diversity from training data manipulation.

Diversity generation from base classifier manipulation:

In this type of method, the diversity of base classifiers are usually generated by two methods: (i) giving different parameters to the base classifier or, (ii) using different types of base classifiers in the ensemble. For example, the decision tree C4.5 of [160] can be run for several times with different parameters values, the ensemble then can be constructed from the diversified decision trees. By using different number of nodes in neural networks, the diversity can be obtained [125]. In [190], seven different types of classifiers were combined for handwritten digits recognition.

Diversity from training data manipulation:

The main method in this category are *feature subset* based techniques. Feature subset based ensemble methods are those that manipulate the input feature set for creating the base classifiers [144, 104, 198, 75]. Some researchers use different partitions of training data, which is capable of producing an ensemble of diverse classifiers [43]. Through randomly partitioning, the original training data can be grouped into some pairwise disjoint subsets, then each base classifier will be trained by an individual subset. Many research results have shown the effectiveness of this approach. Rokach showed that the feature partition is appropriate for classification task with a large number of features [112]. Resampling the original dataset can also be used in Bagging or Boosting [173, 97]. For example, the Attribute Bagging (AB) [157] was proposed by Bryll et al. in 2003. By a random search, AB first finds a suitable size for feature subsets, then the feature subsets are chosen randomly for training base classifiers.

Random Subspace Based Method: Another straightforward strategy to create feature subset based ensemble is random sampling based technique. In this strategy, the feature subsets are obtained by randomly selecting samples from the original training data. Ho proposed the random subspace method in 1998 [78]. A forest of decision trees is produced by pseudo-random selection of subsets from the original training data. Each decision tree is constructed from an individual subset, the forest is obtained

by repeating the constructions of decision trees. Ho showed that the simple random selection of feature subsets is an effective way for constructing ensembles, this is due to the diversity of the base classifiers compensate each other.

The random subspace ensemble method and its variants are widely used in various applications in machine learning and computer vision. A random feature subset based ensemble of Bayesian classifiers was proposed for medical applications [2]. Rodríguez et al. developed the Rotation Forest [164] for classifier ensemble. To create the training data for a base classifier (decision tree), the feature set is randomly split into K subsets (K is a parameter of the algorithm) and Principal Component Analysis (PCA) is applied to each subset. Their experiments showed that rotation forest is more accurate than AdaBoost and Random Forest, and more diverse than these in Bagging, sometimes more accurate as well. In [106], Kuncheva and Rodríguez proposed a Random Linear Oracle ensemble method. Each classifier in the ensemble is replaced by a miniensemble of a pair of subclassifiers with a random linear oracle to choose between the two. It is argued that this approach encourages extra diversity in the ensemble while allowing for high accuracy of the individual ensemble members.

2.2.3 Classifier Output Combination

The methods to combine the base classifiers' outputs can be divided into two classes: weighting and meta-learning [166]. When using weights to combine base classifiers, each base classifier has a proportional contribution to the final decision, the proportion of a classifier is determined by the weight assigned to it. The weight can be fixed or dynamically assigned. Meta-learning is also called as "learning to learn" [136]. If a base learner (classifier) fails to perform efficiently, the meta-learning mechanism itself will adapt in case the same task is presented again.

Weighting Methods

Majority Voting: An instance is assigned to the label which has the highest number of votes from base classifiers in the ensemble. The majority voting can be described by Eqn. (2.1):

$$label(x) = argmax_{c_i \in Y}(\sum_k \delta(y_k(x), c_i)) \tag{2.1}$$

where $y_k(x)$ is the classification result of the k-th base classifier, Y is the domain of $y(x)$, c_i is the label for the i-th class, and $\delta(y, c)$ is the function that:

$$\delta(y, c) = \begin{cases} 1 & \text{if } y(x) = c \\ 0 & \text{if } y(x) \neq c \end{cases} \tag{2.2}$$

Performance Weighting: Using a validation data set, the weights of base classifiers can be tuned based on its classification performance [39]:

$$\alpha(i) = \frac{1 - E_i}{\sum_{j=1}^{T}(1 - E_j)} \tag{2.3}$$

where E_i is a normalization factor which is obtained by the performance of classifier i on a validation data set.

Distribution Summation: The conditional probability vector of each base classifier will be summed up. The instance will be assigned to the class which obtains the highest value from Eqn. (2.4), see [149].

$$label(x) = argmax_{c_i \in Y} \sum_{k} \hat{P}_{M_k}(y = c_i|x) \tag{2.4}$$

where $\hat{P}_{M_k}$ is the probability of x belongs to class c_i produced by classifier M_k.

There is another distribution summation method for posterior probability called *Bayesian Combination*:

$$label(x) = argmax_{c_i \in Y} \sum_{k} P(M_k|S) \cdot \hat{P}_{M_k}(y = c_i|x) \tag{2.5}$$

where $\hat{P}(M_k|S)$ is the probability that classifier M_k is correct, given the training set S.

Vogging: The vogging (Variance Optimized Bagging) method tries to reduce the variance of base classifiers and preserve the pre-defined classification accuracy simultaneously. This is achieved by optimizing a linear combination of base classifiers. The Markowitz Mean-Variance Portfolio Theory is used for obtaining low variance [150].

Meta-learning methods

Stacking: Stacking is a typical meta-learning combination method. It aims to obtain the highest generalization accuracy [70]. The method discriminates base classifiers' reliability by using a meta-learner. The method maintains a meta-dataset, each tuple of this meta-dataset contains the classification predictions from all base classifiers for an instance of the training data. During training, the original data set is partitioned into two subsets, one subset is used to produce the meta-dataset, which is then used for constructing a meta-learner. Another subset is used for constructing base classifiers. A new instance will be first classified by all base classifiers, the predictions will be fed into the meta-leaner to make the final decision.

In [170], the authors pointed out that ensembles with stacking can compete with the best classifier that is selected out from an ensemble by cross-validation. In order to improve the performance of stacking, several variants of the stacking method have been proposed, for example, a weighted combination of stacking and dynamic integration is developed for regression problems [167]. In Troika [132], a new stacking method is proposed by Menahem et al., the new scheme is built from three layers of combined

classifiers. According to the authors, the Troika outperforms traditional stacking, especially in multiclass classification tasks. Jorge et al. proposed the EVOR-STACK [61] method for remote sensing data fusion, the EVOR-STACK uses an evolutionary algorithm for feature weighting, a support vector machine and a weighted kNN stacking is used for classification.

There are other meta-learning combination methods in the literature, for example, grading [176], combiner trees [91] and arbiter trees [90]. However, due to its simplicity and generality, stacking has become a popular selection in meta-learning combination methods.

2.2.4 Ensemble Selection Methods

When constructing an ensemble, one important question is how many base classifiers should be used and which classifiers should be included in the final ensemble. Many researchers insist that a small ensemble can be constructed rather than a larger one, while the classification accuracy and the diverse of the ensemble still can be maintained. The famous "many-could-be-better-than-all" theorem [71] further illustrates that theoretically it is possible to construct small ensembles as strong as the big ones. Ensemble selection is important due to two reasons: efficiency and predictive performance [196]. Ensemble selection has two major approaches: Ranking-based methods and Search-based methods.

Ranking-based Methods

Ranking-base methods set up a criteria to rank the base classifiers, and the classifiers with high ranks will be selected. An agreement-based ensemble selection method was proposed by Margineantu and Dietterich in [38], where the Kappa statistics is used to select pairs of classifiers until the predefined ensemble size is reached. A forward stepwise selection algorithm was proposed in [158], the algorithm selects the classifiers with better performance from thousands of classifiers. Later, a similar algorithm, FS-PP-EROS [153] was proposed by Hu et al., which executes an accuracy-driven forward search to choose the rough subspace classifiers to construct ensemble. Giacinto and Roli developed a dynamic classifier selection (DCS) [62] method to select appropriate classifiers for different instances. In a more recent work, Xiao et al. presented a dynamic classifier ensemble selection method GDES-AD (Group Dynamic Ensemble Selection-Accuracy and Diversity) [208], by using a group method of data handling (GMDH) to DCS, the GDES-AD considers both accuracy and diversity in the process of ensemble selection. The experimental results shown that the GDES-AD has stronger noise-immunity ability than other strategies. Ko et al. proposed a dynamic ensemble selection method [100], the oracle concept was used in their selection scheme, instead of selecting classifiers, their method selects different ensembles for different instances.

Search-based Methods

Being different from ranking, a heuristic search in all possible ensemble subsets is performed using search-based methods. The most representative work in search-based methods is GASEN [71]. GASEN is a selective ensemble method using Genetic Algorithm (GA) to select a subset of neural networks to compose an ensemble, which is better than directly combining all the neural networks. Initially, each neural network is randomly assigned a weight, then GA is used to evolve the weights. After the GA finishes the evolving, the weights will represent classifiers' fitness to join the ensemble. The classifiers have the weights larger than a predefined threshold will be selected into the ensemble. Later, a revised version of GASEN, called GASEN-b [217] was developed to construct ensemble of decision trees, where the weights assigned to classifiers are replaced by bits to indicate their fitness to join the ensemble. In a recent work, a hybrid genetic algorithm (HGA) [98] was proposed for classifier ensemble selection. The HGA is obtained by embedding two local search operations (sequential and combinational) in the standard genetic algorithm. The experiments showed that HGA can obtain better performance than the standard GA.

2.3 Classification with Reject Option and Multi-Stage Classification

In supervised learning, instead of taking a hard decision, allowing for the reject option (no decision made) is of great importance in practice. For instance, in cases of automatic medical diagnosis, it is better to avoid the risk of making a wrong decision when the classifiers cannot make a reliable judgement. Many research results on classification with reject option have shown that a rejection scheme embedded in classification procedure can improve the reliability of classifiers. The multi-stage classification is one of the selection to build a classification system when reject option is employed. When the rejection is not acceptable as a final result, the rejected patterns can be processed at another "higher-stage" pattern recognition system, which would utilize more informative, though more costly measurements [151].

2.3.1 Classification with Reject Option

The theoretical foundation of classification with reject option was built by Chow [29, 30]. In [30], the optimum classification rule with reject option was defined. Suppose $f : \mathbb{R}^k :\rightarrow \{0, 1, R\}$ is a classifier with reject option, which change a binary classification task $Y = \{0, 1\}$ to a three-class situation, where R represents the class of rejection. Denote the probability of assigning an instance x into a class (0 or 1) in Y as $\eta(x)$, the reject probability of f for x is: $p(f(x) = R)$. The misclassification probability is $p(f(x) \neq Y, f(x) \neq R)$. Given a threshold d, the optimum classification rule with reject

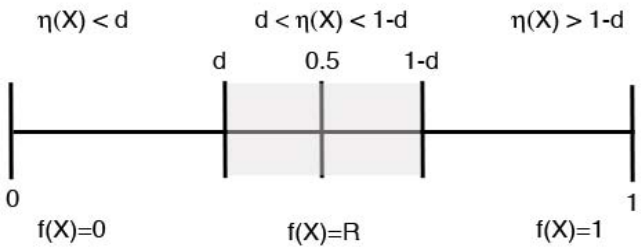


Figure 2.4: Optimum classification rule with threshold d

option can be defined as in Eqn. (2.6) and illustrated in Fig. 2.4:

$$f^{\star}(x) = \begin{cases} 0 & \text{if } 1-\eta(x) > \eta(x) \text{ and } 1-\eta(x) > 1-d \\ 1 & \text{if } \eta(x) > 1-\eta(x) \text{ and } \eta(x) > 1-d \\ R & \text{if } \max(\eta(x), 1-\eta(x)) \leq 1-d \end{cases} \tag{2.6}$$

Chow's rule rejects an instance if its maximum posterior probability is smaller than a predefined threshold. The maximum posterior probability can be used as the reliability measurement of classification. However, it is very hard to get posterior probability in real applications, the posterior probability is often approximated by various types of classifiers such as neural networks [56]. Therefore, finding a reject rule which achieves the best trade-off between error rate and reject rate is undoubtedly of practical interest in real applications.

There are many other rejection rules proposed in literature. Le et al. proposed three different parameters for measuring classification reliability [212]. The most active output, the second most active output and the distance between them are calculated, then three different thresholds are applied on them respectively. Similar rules are also proposed in [122, 183]. A class-relative rejection rule was presented in [59], where the authors suggest that using different rejection thresholds for different classes can obtain better error-reject trade-off. These approaches are proposed to improve the non-optimum estimation of the posterior probability of Chow's rule in real applications. However, the effectiveness of these rules are not theoretically proven. A different type of rejection rule, called *class-selective* rejection was proposed in [195]. Instead of rejecting ambiguous patterns directly, the class-selective remains a list of candidate classes that the pattern more possibly belongs to, i.e. the most possible classes are selected and others are rejected. This is important for some applications such as face recognition, when there is an unrecognized face, people may wish to match it with several possible candidate face images first rather than deny it directly.

Instead of thresholding on the outputs of classifiers, some researchers attempted to embed the reject option into the classifiers, the reject option is determined during classifier training. Most of these attempts focus on support vector machines. In [57], as an extension of SVM, a pair of parallel hyperplanes delimits the rejection region

are provided. The parameters of the hyperplanes can be obtained during the training phase. A similar RO-SVM [222] was proposed by Zhang and Metaxas in 2006, the RO-SVM uses a slight different optimization algorithm to work in the Multiple Instance Learning for image categorization. Bartlett and Wegkamp proposed the optimization of a certain convex loss function φ [8], analogous to the hinge loss used in support vector machines to embed the reject option in SVMs, they showed that minimizing the expected surrogate loss, the φ-risk, also minimizes the risk of misclassification. This work was further extended by Wegkamp in [203] by a generalization of the hinge loss.

The rejection rules for multiple classifier systems have also been proposed in literature. In [54], Foggia et al. proposed to use a unique reliability parameter $\phi \in [0, 1]$ to determine if an instance should be rejected or not. They considered a multiple classifier system where the classifiers are combined using Bayesian rule. Suppose π_1 is the highest estimated posterior probability and π_2 is the second highest one. By combining π_1 and π_2 with appropriate rules to obtain ϕ, the higher value of ϕ indicates more reliable classification. Similar to Chow's rule, a predefined threshold can be used on the value of ϕ to activate rejection. Fumera and Roli analyzed the error-reject trade-off of linearly combined classifiers [58], the conditions under which the weighted average can provide a better error-reject trade-off than the simple average are discussed. When distance-based classifiers are used, or distance-based classifiers and density-based classifiers are combined, their outputs are hard to be compared and combined. Tax and Duin proposed a non-linear transformation *o-norm* for normalizing the outputs of any type of classifiers [193]. In a more recent work [179], the authors studied the possibility to provide ECOC (Error Correcting Output Coding) [44] systems with a tailored reject option carried out through two different schemes: an external and an internal approach. The external approach obtains classification reliability without making any changes on the ECOC system; While in the internal approach, the classification reliability is obtained by estimating the reliability of the internal dichotomizers and implying a slight modification in the decoding stage.

2.3.2 Multistage Classification

When reject option is employed in classification, one has to face the problem of how to deal with the rejected patterns. One way to solve this problem is to pass the rejected patterns to another classifier, which would use more information to treat the rejected patterns. The whole idea of multi-stage classification is to use some more informative measurements by adding them to the set of less informative measurements used in the previous stage. At the final stage a decision is taken in any case, so eventually no rejects remain [151] (Fig. 2.5).

Although the effectiveness of multi-stage classifier was already stressed by some researchers in the late of 1980's and the beginning of 1990's [109, 151], the cascade scheme

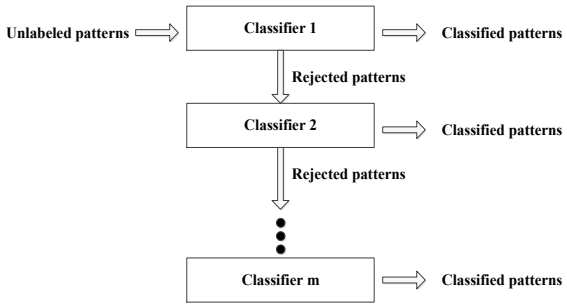


Figure 2.5: A typical multi-stage classification system with m stages

had been almost neglected until the influential work of Viola and Jones [199] published in 2001. The Viola-Jones cascade was developed in the context of face detection, this architecture was used to design the first real time face detector with state-of-the-art accuracy. However, the Viola-Jones architecture does not address the problem of how to automatically determine the optimal cascade configuration, e.g. the numbers of cascade stages and weak learners per stage, or even how to design individual stages so as to guarantee optimality of the cascade as a whole. Therefore, there have been a great number of new cascade systems proposed based on Viola-Jones cascade [123, 171]. Most of these new cascade systems are proposed to solve the problems in face detection [221, 210, 205], object detection [28, 194] and remote sensing image analysis [197, 18].

In many biological and medical applications, people expect high confidence from classifiers. To this end, different cascade schemes can be used to improve the classification reliability. The goal of the first stage is to reduce the number of patterns by rejecting samples with a low confidence. In the following stages, dedicated classifiers are used to determine more difficult patterns. A three-stage classification scheme for ElectroCardioGram (ECG) signal classification was proposed by Hosseini et al. [80]. The first stage is a neural network classifier which detects three types of ECG signals, the signals not in these three classes are rejected and passed into the next stage. The second stage is a similar neural network classifier trained by different types of features, which handles the rejected signals from stage 1. At the last stage, a Self-Organizing Map (SOM) is used to cluster the remaining signals. A similar three-stage framework was proposed by Acır et al. [159] for discriminating electroencephalogram (EEG) signals. A two-stage cascade system for iris image classification was proposed by Sun et al. in 2005 [185]. In order to recognize various iris images efficiently, their proposed cascading scheme uses a local feature classifier (LFC) in the first stage, when the LFC is uncertain of its decision, in the second stage, the LFC and an iris blob matcher are combined to make the final decision. Two cascaded relevance vector machine (RVM) are used in [204] to detect microcalcifications (MC) in digital mammograms. A com-

putationally much simpler linear RVM classifier is applied first to quickly eliminate the overwhelming majority, non-MC pixels in a mammogram from any further consideration. Then another RVM in the second stage is used to determine at each location in the mammogram if an microcalcification object is present or not. The recent applications of reject option and multi-stage classifiers in biomedicine can be found in disease diagnosis [155, 188, 35] and in various types of medical image analysis problems [168, 66, 47, 48, 52].

2.4 One-Class Classification

The term One-Class Classification (OCC) was first proposed by Moya et al. [137], and many approaches have been presented in the literature [192]. Following the taxonomy in the survey papers of [96, 130, 131], the algorithms used in OCC can be categorized as follows: (i) boundary methods, (ii) density estimation and (iii) reconstruction methods.

Tax and Duin tried to separate the positive class from all other patterns in the pattern space; the positive class data was surrounded by a hyper-sphere which encompassed almost all positive patterns within the minimum radius [189, 191]. Their method of Support Vector Data Description (SVDD) was different with that proposed by Schölkopf et al. [175] who used a separating hyper-plane instead of a hyper-sphere to separate the pattern space with data from the space containing no data. Manevitz and Yousef [129] proposed another version of one-class SVM to identify the outlier data as representative of the second class with the standard *Reuters*[1] dataset. They noted that their SVM methods was quite sensitive to the choice of representation and kernel. Although one-class classifiers, such as OCSVM, have been widely used, the estimated boundary can be sensitive to the nature of the data [169]. When noisy data, or many outliers, are contained in the training set, OCSVM will generate a large boundary that encloses regions of the feature space where the positive class has low density, often resulting in many false positives [79]. This can be highly problematic for many applications, especially for medical diagnosis where the percentage of outliers must be kept to a minimum, since an accidental diagnosis of a patient as healthy may result in serious consequences.

Density estimation methods estimate the density of the target class to form a model to represent the data. The generally used models include Parzen, Gaussian and Gaussian mixture models. A test point is classified by the maximum posterior probability. Density estimation methods work well if the number of training samples is sufficient enough to estimate data distributions. However, when the models cannot fit the data distribution very well, a large bias may be generated. Details and some comparisons of these methods can be found in [162, 202].

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578

When it is not feasible to obtain large training sets, which are required by the density estimation or support vector based methods, the reconstruction models can be used to approximate the target class. The reconstruction models aim to produce prototypes of the original data, and new objects are projected onto the prototypes. The distance between the original object x and the projected object $p(x)$ (Reconstruction Error) indicates the similarity of a new object to the original target distribution. When the training data has a very high dimensionality, some distance based methods like nearest neighbor tend to perform poorly [12]. In such cases it can often be assumed that the target data is distributed in subspaces of much lower dimensionality. Principle Component Analysis [186] is a linear model that has the ability to project the original data into orthogonal space which can capture the variance in the data. In order to approximate nonlinear data distributions, many nonlinear subspace models have also been proposed, such as Self-Organizing Map (SOM), auto-encoders, auto-associative networks and Kernel PCA.

It has been demonstrated that combining classifiers can also be effective for one-class classifiers. The existing classifier combination strategies can be used in one-class classifiers. However, since there is only information from one class, it is more difficult to combine one-class classifiers. Tax and Duin investigated the influence of feature sets and the types of one-class classifiers for the best choice of the combination rule [190]. A bagging based one-class support vector machine ensemble method was proposed in [178]. A dynamic ensemble strategy based on Structural Risk Minimization [86] was proposed by Goh et al. for multiclass image annotation [65]. Recently, some research results have revealed that creating a one-class classifier ensemble from different feature subsets can provide better performance. Perdisci et al. [152] also used an ensemble of one-class SVMs to create a "high speed payload-based" anomaly detection system, the features were first extracted and clustered, the OCSVM ensemble was then constructed based on the clustered feature subsets. A biometric classification system combining different biometric features was proposed by Bergamini et al. [10], where the one-class SVMs in the ensemble were trained by the data from different people. The feature subset strategy provides diversity with respect to the base classifiers, some researchers emphasis the importance of measuring diversity in ensembles so as to improve classification performance [72, 102].

Combining one-class classifiers has also shown promising performance in medicine and biology [213]. Peng Li et al. [116] proposed a multi-size patch-based classifier ensemble, which provides a multiple-level representation of image content, the proposed method was evaluated on colonoscopy images and ECG beat detection [115]. The k-nearest neighbor classifier was selected as the base classifier in the work of Okun and Priisalu [146]; majority voting was chosen as the combination rules for the ensemble; the method was evaluated on gene expression cancer data.

Chapter 3

Random Subspace Ensemble of Neural Networks for Microscope Image Classification

The content of this chapter has been published in the following papers:

- Yungang Zhang, Bailing Zhang and Wenjin Lu. Breast Cancer Classification From Histological Images with Multiple Features and Random Subspace Classifier Ensemble, CMLS 2011, AIP Conf. Proc. Vol. 1371, pp. 19-28, Toyama, Japan, June 2011.
- Yungang Zhang, Bailing Zhang and Wenjin Lu. Breast Cancer Histological Image Classification with Multiple Features and Random Subspace Classifier Ensemble, T.D. Pham, L.C. Jain (eds): Innovations in Knowledge-based Systems in Biomedicine, Springer-Verlag, SCI 450, pp. 27-42, 2013. (book chapter).
- Bailing Zhang, Yungang Zhang, Wenjin Lu and Guoxia Han. Phenotype Recognition by Curvelet Transform and Random Subspace Ensemble. Journal of Applied Mathematics & Bioinformatics, Vol.1, No.1, pp. 79-103, 2011.

3.1 Introduction

Automated microscopic image analysis has become a fundamental tool for scientists to make discovery in biological and medical science. Modern robotic fluorescence microscopes are able to capture thousands of images from massively parallel experiments such as RNA interference (RNAi) or small-molecule screens. High-content screening has become a drug discovery method that uses images of living cells as the basic unit for molecule discovery, which permits the identification of small compounds altering cellular phenotypes. As such, efficient computational methods are required for automatic cellular phenotype identification capable of dealing with large image data sets.

The classification or identification of cellular phenotype is often a rate limiting task because of the high dimensionality and small sample size of the microscopy images.

Complex cellular structures such as organelles within the eukaryotic cell can be studied by fluorescence microscopy images of cells with appropriate staining techniques. By robotic systems, thousands of images from cell assays can be acquired from the so-called High-Content Screening (HCS), which often yields high-quality, biologically relevant information. Many biological properties of the cell can be further analyzed from the images, for example, the size and shape of a cell, amount of fluorescent label, DNA content, cell cycle, and cell morphology [84]. On the other hand, High-Throughput Screening or HTS allows a researcher to quickly conduct millions of biochemical, genetic or pharmacological tests using robotics, data processing and control software, liquid handling devices, and sensitive detectors. The high-content, high-throughput screening has greatly advanced biologists' understanding of complex cellular processes and genetic functions [124]. With the aid of computer vision and machine learning, scientists are now able to carry out large-scale screening of cellular phenotypes, at whole-cell or sub-cellular levels, which are important in many applications, e.g., delineating cellular pathways, drug target validation and even cancer diagnosis [216, 88].

The high-content screening has also significantly facilitated genome-wide genetic studies in mammalian cells. With the combination with RNA interference (RNAi), sets of genes involved in specific mechanisms, for example cell division, can be identified. By observing the downstream effect of perturbing gene expression, genes' normal operations that function to produce proteins needed by the cell can thus be assessed [138]. RNAi is a phenomenon of degrading the complementary mRNA by introduction of double-stranded RNA (dsRNA) into a diverse range of organisms and cell types [87, 64]. The discovery of RNAi and the availability of whole genome sequences allow the systematic knockdown of every gene or specific gene sets in a genome [31] . Libraries of RNAis, covering a whole set of predicted genes inside the target organisms genome can be used to identify relevant subsets, facilitating the annotation of genes for which no clear role has been established beforehand. Image-based screening of the entire genome for specific cellular functions thus becomes feasible by the development of Drosophila RNAi technology to systematically disrupt gene expression. Genome-wide screens, however, produce huge volumes of image data which is beyond human's capability of manual analysis, and automating the analysis of the large number of images generated in such screens is the bottleneck in realizing the full potential of cellular and molecular imaging studies.

Microscope imaging is also an important tool in the diagnosis of many types of diseases. For example, histopathologic biopsy images are widely accepted as a powerful gold standard for prognosis in critical diseases such as breast, prostate, kidney and lung cancers, allowing to narrow borderline diagnosis issued from standard macroscopic

non-invasive analysis such as mammography and ultrasonography [83], and histopathology slides provide a comprehensive view of disease and its effect on tissues, since the preparation process preserves the underlying tissue architecture [68].

In this chapter, an approach based on multiple image feature descriptions and random subspace ensemble for microscopic image classification is investigated. Three types of image feature descriptors are used for microscopic image description: the curvelet transform, the gray level co-occurrence matrix (GLCM), and the completed local binary patterns (CLBP). The curvelet transform is a multiscale directional transform which allows an almost optimal nonadaptive sparse representation of objects with edges [20], which is particularly appropriate for many microscopy images. The GLCM and CLBP give the textural descriptions for the microscopic images. The ensemble classification approach, called Random Subspace Ensemble, contains a set of base neural network classifiers which are trained using subsets of curvelet features randomly drawn from the available RNAi images. The component classifiers are then selected and aggregated by following the Majority Voting Rule. Experimental results on the phenotype recognition from three benchmark fluorescence microscopy image sets(RNAi, CHO and 2D-Hela) and a breast cancer biopsy image set show the effectiveness of the proposed approach. The ensemble model produces better performance compared to any of individual neural networks trained. The proposed Random Subspace Ensemble offers the classification rate 87.4% on the RNAi image dataset, which compares sharply with the published result 82%, and the classification results on the other two groups of fluorescence microscopy images (CHO and 2D-Hela) certify the effectiveness of the proposed approach as well. The performance of the proposed classification method also superior than the published results on a breast cancer biopsy image set.

The chapter is organized as follows: In Section 3.2, some related works are presented. Section 3.3 introduces the image data used. The image feature extraction methods are described in Section 3.4 and the random subspace ensemble of neural networks is introduced in Section 3.5. Section 3.6 presents the experimental results, the conclusion is drawn in Section 3.7.

3.2 Related Works

Most of the microscopic image analysis systems consist of several components: cellular segmentation, cellular morphology and texture feature extraction, cellular phenotype classification, and clustering analysis [84]. With appropriate cellular segmentation results, phenotype recognition can be studied in a multi-class classification framework, which involves two interweaved components: feature representation and classification. Efficient and discriminative image representation is a fundamental issue in any bioimage recognition task. Most of the proposed approaches for microscopic images employ feature set which consist of different combinations of morphological, edge, texture, ge-

ometric, moment and wavelet features [113], and most of these systems employ SVM or neural networks as their classifiers, graphical models are also used for classification [154, 27].

In a early work of Boland et al. in 1998 [139], the Zernike moments and Haralick texture features are combined for fluorescence microscope image classification, using a backpropagation neural network, an averaged accuracy of 88% was obtained on a Chinese Hamster Ovary (CHO) cells images. In 2001, Boland and Murphy further extended their feature description method in [139], besides Zernike moments and Haralick texture features, the Subcellular Location Features (SLF) was first proposed for microscope image classification [140], which is a combination of features from the whole image and cellular structures. Using a neural network classifier, they obtained 83% classification accuracy on a 10-class microscope image set 'Hela'. The author also demonstrated that SLF have better description ability than other frequently used features like Zernike moments and Haralick texture features. Zhao et al. used clustering methods for object type recognition in micrscopic images, k-means is first used to cluster the subcellular location patterns, then a linear discriminant analysis (LDA) classifier is employed to discriminate different objects in subcellular location patterns, 83% recognition rate was obtained. In these works, SLF achieves better performance than other image features. However, in SLF, some features cannot be obtained without segmentation of images, this makes SLF not suitable for all types of microscopic images.

The classification of microscopic images without image segmentation has been addressed by many researchers. In 2004, Huang et al. proposed to use only the global features from SLF to solve the 'type-specific' problem in SLF [92], the stepwise discriminant analysis is used to select the most discriminative features, better classification results were obtained on multicell images than single-cell images. The authors further improved the performance of their non-segmentation method by adding Gabor and wavelets features and ensemble learning [93]. The 13 statistic features from co-occurrence matrix are used in [81, 82] for time series microscopic image classification, the features are calculated for all images, then the mean and variance across the series are used as the final features of the images. Recently, some researchers have utilized multiresolution image features in microscopic image classification [1, 46, 14, 187]. These works have showed that the multiresolution features are very suitable for describing subtle structures in microscopic images.

3.3 Microscope Image Data

Three benchmark fluorescence microscopy image datasets in [113] were used in our study, which are RNAi, CHO and 2D-Hela. A benchmark breast cancer biopsy image dataset was also used for evaluation.

3.3.1 Fluorescence microscope image data

The RNAi dataset is a set of fluorescence microscopy images of fly cells (D. melanogaster) subjected to a set of gene-knockdowns using RNAi. The cells are stained with DAPI to visualize their nuclei. Each class contains 20 1024×1024 images of the phenotypes resulting from knockdown of a particular gene. Ten genes were selected, and their gene IDs are used as class names. The genes are CG1258, CG3733, CG3938, CG7922, CG8114, CG8222, CG 9484, CG10873, CG12284, CG17161. According to [113], the images were acquired automatically using a Delta-Vision light microscope with a 609 objective. Each image is produced by deconvolution, followed by maximum intensity projection (MIP) of a stack of 11 images at different focal planes. Samples of the images are illustrated in Fig. 3.1.

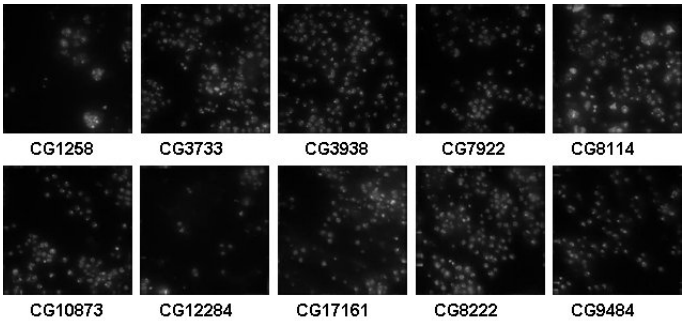


Figure 3.1: RNAi image set of fluorescence microscopy images of fly cells (D. melanogaster).

2D HeLa dataset, a collection of HeLa cell immunofluorescence images containing 10 distinct subcellular location patterns. The subcellular location patterns in these collections include endoplasmic reticulum (ER), the Golgi complex, lysosomes, mitochondria, nucleoli, actin microfilaments, endosomes, microtubules, and nuclear DNA (Fig. 3.2). The 2D HeLa image dataset is composed of 862 single-cell images, each with size 382×512.

CHO is a dataset of fluorescence microscope images of CHO (Chinese Hamster Ovary) cells. The images were taken using 5 different labels. The labels are: anti-giantin, Hoechst 33258 (DNA), anti-lamp2, anti-nop4, and anti-tubulin (Fig. 3.3). The CHO dataset is composed of 340 images, each with size 512×382.

3.3.2 Breast Cancer Biopsy Image Set

We used a breast cancer benchmark biopsy image dataset from the Israel Institute of Technology [1]. 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

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

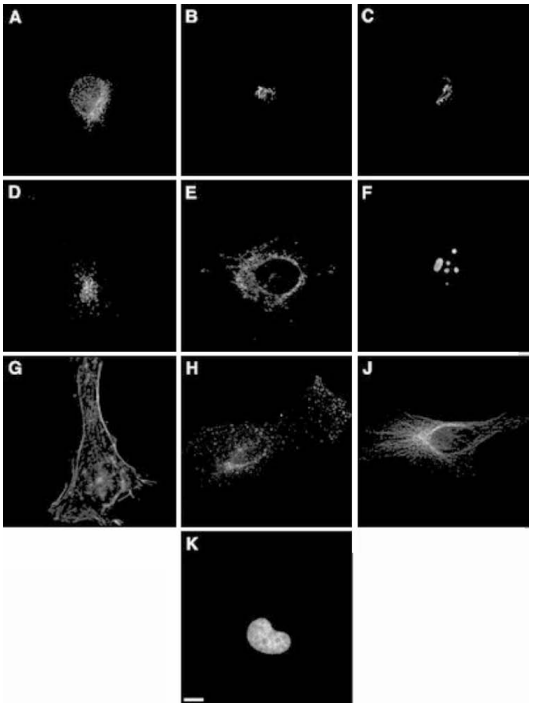


Figure 3.2: Representative images from the 2-D HeLa image collection. The image classes represent the distributions of (a) an endoplasmic reticulum (ER) protein, (b) the Golgi protein giantin, (c) the Golgi protein GPP130, (d) the lysosomal protein LAMP2, (e) a mitochondrial protein, (f) the nucleolar protein nucleolin, (g) the filamentous form of the cytoskeletal protein actin, (h) the endosomal protein transferrin receptor, (j) the cytoskeletal protein tubulin, and (k) the fluorescent probe DAPI bound to DNA [113].

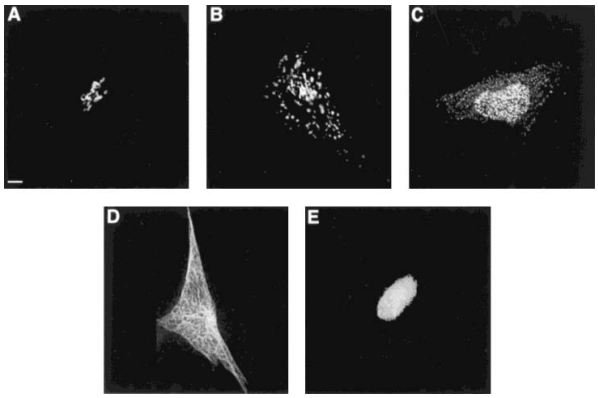


Figure 3.3: Examples of the images in CHO dataset. These images have had background fluorescence subtracted and have had all pixels below threshold set to 0. Representative images are shown for cells labeled with antibodies against giantin (A), LAMP2 (B), NOP4 (C), tubulin (D), and with the DNA stain Hoechst 33258 (E) [113].

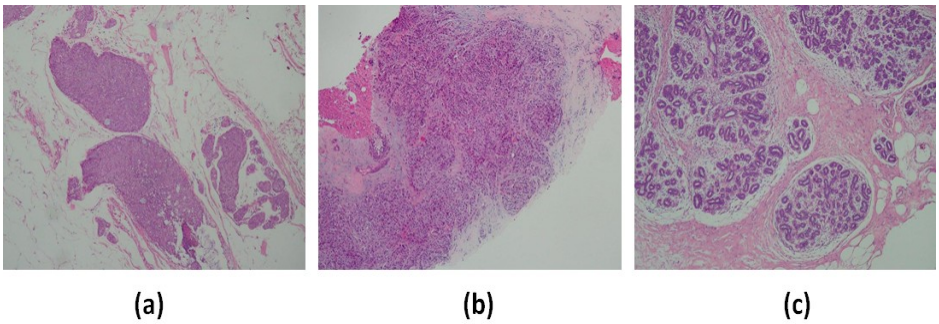


Figure 3.4: (a) carcinoma in situ: tumor confined to a well-defined small region; (b) invasive: breast tissue completely replaced by the tumor; (c): healthy breast tissue.

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 5μ 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. Three typical sample images belong to different classes can be seen in Fig. 3.4.

3.4 Feature Extraction

Shape feature and texture feature are critical factors for distinguishing one image from another. For the microscopic image discrimination, shapes and textures are also quite effective. As we can see from Fig. 3.1 to Fig. 3.4, different kinds of microscope 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 multiresolution shape description.

3.4.1 Curvelet Transform for Image Feature Description

Although wavelets have been widely used in image analysis, traditional wavelets perform well only at representing point singularities, since they ignore the geometric properties of structures and do not exploit the regularity of edges. Curvelet transform was proposed in order to overcome the drawbacks of conventional wavelet transform, the curvelet transform has an almost optimal sparse representation of objects with C^2–singularities [24], combined with other methods, superior performance of the curvelet transform has been shown in image processing [128].

The Continous Curvelet Transform

In this section we briefly introduce the continuous curvelet transform (CCT) in [21, 22]. Curvelets functions can be constructed from two window functions $V(t)$ and $W(r)$ (for example, the scaled Meyer windows [42]), which satisfy the following admissibility conditions:

$$\sum_{l=-\infty}^{\infty} V^2(t-l) = 1, \quad t \in \mathbb{R}, \tag{3.1}$$

$$\sum_{j=-\infty}^{\infty} W^2(2^j r) = 1, \quad r > 0, \tag{3.2}$$

$$\int_0^{\infty} W^2(r)\frac{dr}{r} = \ln 2, \tag{3.3}$$

$$\int_{-1}^{1} V^2(t)dt = 1. \tag{3.4}$$

With three parameters, the *scale* $a \in (0, 1]$, the *location* $b \in \mathbb{R}^2$ and the *orientation* $\theta \in [0, 2\pi)$, using the polar coordinates(r, ω) in frequency domain, the a-scaled window can be defined as:

$$U_a(r, \omega) := a^{\frac{3}{4}} W(ar) V(\frac{\omega}{\sqrt{a}}) \tag{3.5}$$

Let the Fourier transform for a function $f \in L^2(\mathbb{R}^2)$ be defined by:

$$\hat{f}(\xi) := \frac{1}{2\pi} \int_{\mathbb{R}^2} f(x) e^{-i\langle x,\xi\rangle} dx. \tag{3.6}$$

Designate the window U_a as the Fourier transform of the curvelet function $\Phi_{a,0,0}$, we can get:

$$\hat{\Phi}_{a,0,0}(\xi) := U_a(\xi). \tag{3.7}$$

The curvelet family can be constructed by translation and rotation of $\Phi_{a,0,0}$,

$$\Phi_{a,b,\theta} := \Phi_{a,0,0}\big(R_\theta(x-b)\big), \tag{3.8}$$

where the translation $b \in \mathbb{R}^2$ and $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ is the rotation matrix with angle θ. Fig. 3.5 is an example graph of a curvelet function.

The *continuous curvelet transform* Γ_f of the function $f \in L^2(\mathbb{R}^2)$ is given as:

$$\Gamma_f(a, b, \theta) := \langle \Phi_{a,b,\theta}, f \rangle = \int_{\mathbb{R}^2} \Phi_{a,b,\theta}(x)\overline{f(x)}dx, \tag{3.9}$$

Γ_f is the product of a given function f with every curvelet element $\Phi_{a,b,\theta}$.

Figure 3.5: Graph of a curvelet function with $\Phi_{a,b,\theta}$, $a = 2^{10}, b = 0, \theta = 120\,^\circ$

The Discrete Curvelet Transform

It is necessary to discretize the continuous curvelet transform, since we usually work with discrete data. The idea of discretizing the continuous curvelet transform is simple—choose a suitable sampling at the range of scales, locations and directions. The scales $a_j := 2^{-j}, \quad j \geq 0$; the equidistant sequence of rotation angles $\theta_{j,l}$:

$$\theta_{j,l} := \frac{\pi l 2^{-\lceil j/2 \rceil}}{2}, \quad l = 0, 1, \ldots, 4 \cdot 2^{\lceil j/2 \rceil} - 1; \tag{3.10}$$

the positions: $b_k^{j,l} = b_{k_1,k_2}^{j,l} := R_{\theta_{j,l}}^{-1}(\frac{k_1}{2^j}, \frac{k_2}{2^{j/2}})^T$, with $k_1, k_2 \in \mathbb{Z}$ and R_θ denotes the rotation matrix with angle θ.

This choice leads a discrete curvelet transform (DCT) forms a tight frame, hence the discrete curvelet transform will be invertible. The choice of positions yields a parabolic scaling of the grids with the relationship length$\approx 2^{-j/2}$ and width$\approx 2^{-j}$ (Fig. 3.6).

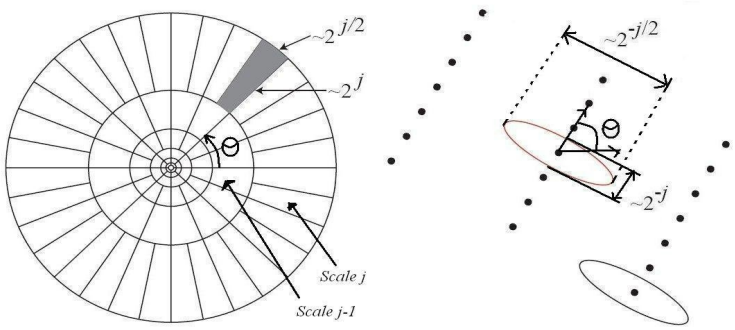


Figure 3.6: Curvelet transform: Fourier frequency domain partitioning (left) and spatial domain representation of a wedge (right)

However, in practical implementation, we prefer Cartesian arrays to the polar tiling of the frequency plane, therefore, a construction of coronae based on concentric squares

and shears can be seen in Fig. 3.7 [25].

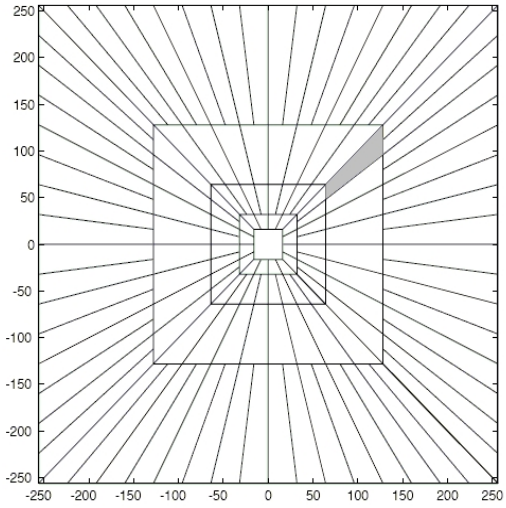


Figure 3.7: Discrete curvelet tiling coronae

We demonstrate DCT applied to an image given by $f(x_1, x_2)$,$x_1 = 0, 1, \ldots, N_1 - 1$,$x_2 = 0, 1, \ldots, N_2 - 1$, whose discrete Fourier transform is

$$\hat{f}(n_1, n_2) = \sum_{x_1=0}^{N_1-1} \sum_{x_2=0}^{N_2-1} f(x_1, x_2) e^{-2\pi i(n_1 x_1/N_1 + n_2 x_2/N_2)}. \tag{3.11}$$

The discrete curvelet transform Φ_{jlk} decomposites the image f into the curvelet coefficients c_{jlk},

$$f(x_1, x_2) = \sum_{j=1}^{J} \sum_{l=0}^{L_j-1} \sum_{k_1=0}^{K_{jl,1}-1} \sum_{k_2=0}^{K_{jl,2}-1} c_{jlk} \Phi_{jlk}(x_1, x_2), \tag{3.12}$$

where $k = (k_1, k_2)$ and Φ_{jlk} is the curvelet on level j with orientation l and spatial translation k. The discrete curvelet transform thus provides a decomposition of the image f into J detail sub-bands(scales), with L_j directions on each level, and $K_{jl,1} \times K_{jl,2}$ spatial translations for each of these directions [23].

The *discrete curvelet transform* can be defined through its discrete Fourier transform as

$$\hat{\Phi}_{j0k}(n_1, n_2) = U_j(n_1, n_2) e^{-2\pi i(k_1 n_1/K_{j0,1} + k_2 n_2/K_{j0,2})} \tag{3.13}$$

and

$$\hat{\Phi}_{jlk} = S_{\theta_l}^T \hat{\Phi}_{j0k}. \tag{3.14}$$

S_{θ_l} is a shearing matrix, which shears the grid on which the curvelet is evaluated by an angle θ_l. The slopes defined by the angles θ_l are equispaced. U_j is a frequency window function with compact support and defined as

$$\sum_j \sum_l \left| [S_{\theta_l} U_j](n_1, n_2) \right|^2 = 1. \tag{3.15}$$

Therefore, the discrete curvelet transform decomposites the frequency space into dyadic rectangular coronae, each of which is divided into wedges, the number of wedges doubles with every second level. This is how the frequency coronae in Fig. 3.7 be constructed.

Two different digital (or discrete) curvelet transform (DCT) algorithms are introduced in [25]. The first algorithm is the Unequispaced FFT Transform, where the curvelet coefficients are found by irregularly sampling the fourier coefficients of an image. The second algorithm is the the Wrapping transform, using a series of translations and a wraparound technique. Both algorithms having the same output, but the Wrapping Algorithm gives both a more intuitive algorithm and faster computation time. In this thesis the Wrapping DCT method is used. The Wrapping DCT algorithm can be briefly described as follows:

1. Take FFT(Fast Fourier Transform) of the image.
2. Divide FFT into collection of Digital Corona Tiles (Fig. 3.7).
3. For each corona tile

(a) Translate the tile to the origin.

(b) Wrap the parallelogram shaped support of the tile around a rectangle centered at the origin.

(c) Take the Inverse FFT of the wrapped support.

(d) Add the curvelet array to the collection of curvelet coefficients.

The inverse Wrapping DCT algorithm is:

1. For each curvelet coefficient array

(a) Take the FFT of the array.

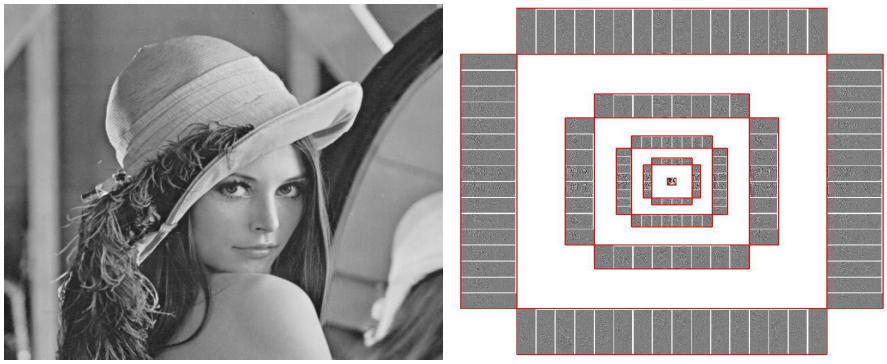
(b) Unwrap the rectangular support to the original orientation shape.

(c) Translate to the original position.

(d) Store the translated array.

2. Add all the translated curvelet arrays.
3. Take the inverse FFT to reconstruct the image.

The details of the algorithms can be found in [25]. Fig. 3.8(b) shows the curvelet coefficients of a 6-level decomposition of a 512×512 *Lena* in Fig. 3.8(a). On the coarsest level, $j = 1$, the curvelets are isotropic, the low-pass image is located at the center of the coronae, the sub-bands curvelet coefficients located around the low-pass image according to their scales and orientations, and on the finest level, $j = J$ ($j = 6$ in Fig. 3.8(b)), one can choose to use curvelet or wavelet in the implementation, we have used wavelets on the finest level since the shorter execution time and smaller memory requirements. Actually, there is a rule to determine the decomposition levels according to the size of the image, the number of decomposition levels DL can be calculated as $DL = \log 2(n) - 3$, where $n = min(M, N)$ for a $M \times N$ size image.



(a) Lena (b) Decomposition Coefficients

Figure 3.8: 6-level DCT decomposition

Multiresolution Feature Extraction Through Curvelet Transform

Fig. 3.9 gives an example of applying curvelet transform on a microscopic image. The top image is the original one. The first image in second row is the approximate coefficients and others are detailed coefficients at eight angles from three scales. All the images are rescaled to same dimension for demonstration purpose.

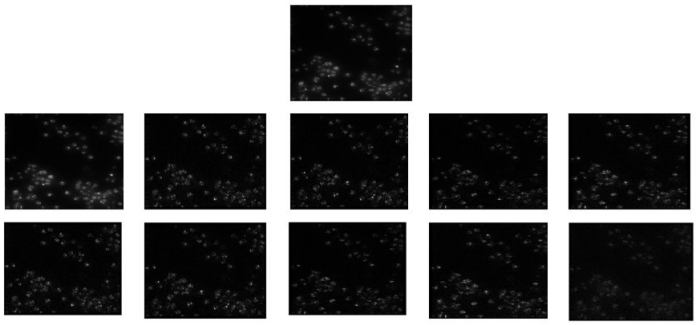


Figure 3.9: Curvelet transform of a RNAi miscroscopy image

Once the curvelet coefficients have been obtained from DCT, the mean values and standard deviations of each curvelet sub-band are computed as the features for the curvelet, since these features have shown good capability in description of wavelet and curvelet sub-bands [7, 184]. If n curvelets are used for the transform, $2n$ features $G = [G_\mu, G_\delta]$ are obtained, where $G_\mu = [\mu_1, \mu_2, \cdots, \mu_n]$, $G_\delta = [\delta_1, \delta_2, \cdots, \delta_n]$. The $2n$ dimension feature vector can be used to represent each image in the dataset.

3.4.2 Completed Local Binary Patterns for Texture Description

Local Binary Patterns (LBPs) were first introduced as a texture descriptor for summarizing local gray-level structures [145], 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

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \quad s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{3.16}$$

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 [145]. 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 2 transitions) is a uniform pattern, whereas 11110101 (with 4 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. [145] observed that although only 58 of the 256 8-bit patterns are uniform, nearly 90 percent of all observed image neighborhoods are uniform. We use the notation $\text{LBP}^u_{P,R}$ 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 $\text{LBP}^u_{8,1}$.

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 [67]. 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 $CLBP\_C$ (complete local binary patterns of centers). Two other complementary components are also obtained

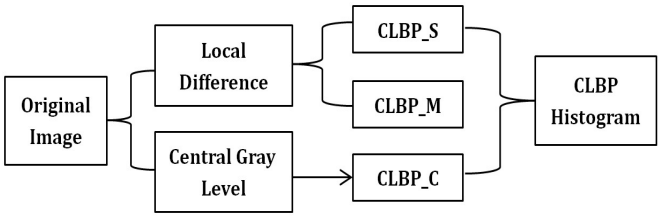


Figure 3.10: Framework of CLBP

by LDSMT: the difference signs and the difference magnitudes, two operators $CLBP\_S$ (complete local binary patterns of signs) and $CLBP\_M$ (complete local binary patterns of magnitudes) are used to code them. The framework of CLBP is presented in Fig. 3.10. The CLBP could achieve much better rotation invariant texture classification than conventional LBP based schemes.

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:

$$d_p = s_p * m_p, \quad and \begin{cases} s_p = sign(d_p) \\ m_p = |d_p| \end{cases} \tag{3.17}$$

where $s_p = 1$, when $d_p \geq 0$, otherwise, $s_p = 0$. m_p is the magnitude of d_p. Eqn. 3.17 is called the local difference sign-magnitude transform (LDSMT).

The $CLBP\_S$ operator is defined as the original LBP operator in Eqn. 3.10.

The $CLBP\_M$ operator is defined as:

$$CLBP\_M_{P,R} = \sum_{p=0}^{P-1} t(m_p, c)2^p, \quad t(x, c) = \begin{cases} 1 & \text{if } x \geq c \\ 0 & \text{if } x < c \end{cases} \tag{3.18}$$

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

The $CLBP\_C$ operator is coded as:

$$CLBP\_C_{P,R} = t(g_c, c_I) \tag{3.19}$$

where t is defined in Eq. 3.18 and c_I is a threshold set as the average gray level of the whole image.

In this work, we have used the 3D joint histogram of these three operators to generate textural features of breast cancer biopsy images, according to [67], the joint combination of the three components gives better classification than conventional LBP and provides a smaller feature dimension.

Table 3.1: 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

Statistics from Gray Level Co-occurrence Matrix

Global texture distribution is one of the important characteristics used for image description. The co-occurrence probabilities provide a second-order statistics for generating texture features [76]. The basis for features used here is the gray level co-occurrence matrix, which 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 [13], the matrix can be seen in Eqn. 3.20.

$$\mathbf{C} = \begin{bmatrix} p(1,1) & p(1,2) & \cdots & p(1,N_g) \\ p(2,1) & p(2,2) & \cdots & p(2,N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g,1) & p(N_g,2) & \cdots & p(N_g,N_g) \end{bmatrix} \tag{3.20}$$

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 3.1. 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 [76, 13, 32, 182].

3.4.3 The 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 higher clas-

sification accuracy. Using five levels of the curvelet transform, 82 sub-bands of curvelet coefficients are computed, therefore, a 164 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 3.1 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 584 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 Section 3.5) this problem can be resolved due to its dimension reduction capability.

3.5 Random Subspace Ensemble of Neural Networks

The idea of classifier ensemble is to individually train a set of classifiers and appropriately combine their decisions [108]. The variance and bias of classification can be reduced simultaneously because the collective results will be less dependent on peculiarities of a single training set while a combination of multiple classifiers may learn a more expressive concept class than a single classifier. Classifier ensembles generally offer improved performance. There are many ways to form a classifier ensemble. A mainstream methodology is to train the ensemble members on different subsets of the training data, which can be implemented by re-sampling (bagging) [111] and re-weighing (boosting) [211] the available training data. Bagging (an abbreviation of "bootstrap aggregation") uses the bootstrap, a popular statistical re-sampling technique, to generate multiple training sets and to train base classifiers for an ensemble. Boosting generates a series of component classifiers whose training sets are determined by the performance of former ones. Training instances that are wrongly classified by the former classifiers will play more important roles in the training of later classifiers.

Though different classifiers can be applied in ensemble learning, in this chapter we will mainly consider neural classifiers as the base learners with the following reasons. First of all, it has been proven that a simple three-layer back propagation neural network (BPNN) can approximate any continuous function if there are sufficient number of middle-layer units [77]. Secondly, the generalization performance of neural networks is not very stable in the sense that different settings such as different network architectures and initial conditions may all influence the learning outcome. The existence of such differences between base classifiers is pre-requisite for the success of a classifier ensemble [108].

The multilayer perceptron (MLP) trained with the back propagation algorithm has been successfully applied to many classification problems in bioinformatics, for exam-

ple, subcellular protein location patterns [139, 93, 140]. With a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and a layer of output nodes, an MLP constructs input–output mappings and the characteristics of such input–output relationship are determined by the weights assigned to the connections between the nodes in the two adjacent layers. Changing the weight will change the input-to-output behavior of the network. An MLP learning or training is often implemented by gradient descent based back-propagation algorithm [77] to optimize a derivable criterion, such as the Mean Squared Error.

The performance improvement can be expected from an MLP ensemble by taking advantages of the disagreements among a set of MLP classifiers. An important issue in constructing the MLP ensemble is to create the diversity of the ensemble [78]. The main idea of Random Subspace is: for a p-dimensional training set, choose a fixed p^* ($p^* < p$), randomly select p^* features according to the uniform distribution. Thus, the data of the original p-dimensional training set is transformed to the selected p^*-dimensional subspace. The resulting feature subset is then used to train a suitable base classifier. Repeat this process for m times, then m base classifiers are trained on different randomly chosen feature subsets, the resulting set of classifiers are then combined by majority voting. Random Subspace simultaneously encourages diversity and individual accuracy within the ensemble: random feature sets selection results in diversity among the base classifiers and using the corresponding data set to train each base classifier prompt the accuracy. The details of Random Subspace Ensemble can be further described as follows:

Consider a training set $X = \{X_1, X_2, \ldots, X_n\}$ with n samples, each sample is assigned into one of m classes, $m \geq 2$. Each training sample X_i is described by a p-dimensional vector, $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}(i = 1, \ldots, n)$. We randomly select $p^* < p$ features from the original p-dimensional feature vector to obtain a new p^*-dimensional feature vector. Now the original training sample set X is modified as $X^r = \{X_1^r, X_2^r, \ldots, X_n^r\}$, each training sample in X^r is described by a p^* feature vector, $X_i^r = \{x_{i1}^r, x_{i2}^r, \ldots, x_{ip^*}^r\}(i = 1, \ldots, n)$, where each feature component $x_{ij}^r (j = 1, \ldots, p^*)$ is randomly selected according to the uniform distribution. Then we construct R classifiers in the random subspace X^r and aggregate these classifiers in the final majority voting rule. This procedure can be formally described as:

1. Training phase. Repeat for $r = 1, 2, \ldots, R$.

   (a) Select the p^*-dimensional random subspace X^r from the original p-dimensional feature space X. Denote each p^*-dimensional feature vector by x.

   (b) Construct a classifier $C^r(x)$ (with a decision boundary $C^r(x) = 0$) in X^r.

2. Classification phase. Combine classifiers $C^r(x), r = 1, \ldots, R$ by majority voting to a final decision rule $\beta(x) = argmax_{y \in \{1,\ldots,m\}} \sum_r \delta_{sgn(C^r(x)),y}$, where $\delta_{i,j} = 1$,

if $i = j$. Otherwise, $\delta_{i,j} = 0$. $y \in \{1, \ldots, m\}$ is a decision (class label) of the classifier.

3.6 Experiments and Results

For the curvelet feature extraction process, fast discrete curvelet transform was applied to each of the images in the database using the CurveLab Toolbox (http://www.curvelet.org). By following the four steps described in Section 3.4.1: application of a 2-dimensional FFT of the image, formation of a product of scale and angle windows, wrapping this product around the origin, and application of a 2-dimensional inverse FFT. The discrete curvelet transform can be calculated to various resolutions or scales and angles. Two parameters are involved in the digital implementation of the curvelet transform: number of resolutions and number of angles at the coarsest level. For our images of 1024×1024, five scales were chosen which include the coarsest wavelet level. At the 2nd coarsest level 16 angles were used. With five levels analysis, 82 sub-bands of curvelet coefficients are computed. Therefore, a 164 dimension feature vector is generated for each image in the database.

We first evaluated several different and commonly used supervised learning methods to the multi-class classification problem, including k-nearest neighbors (kNN), multi-layer perceptron neural networks, SVM, Random Forest and random subsapce ensemble. kNN classifier is prototype-based, with an appropriate distance function for comparing pairs of data samples. It classifies a sample by first finding the k closest samples in the training set, and then predicting the class by majority voting. We simply chosen $k = 1$ in the comparisons. Multiple layer perceptron (MLP) is configured as a structure with one hidden layer with a few hidden units. The activation functions for hidden and output nodes are logistic sigmoid function and linear function, respectively. We experimented with MLP with 20 units in the hidden layer and 10 linear units representing the class labels. The network is trained using the Conjugate Gradient learning algorithm for 500 epochs.

Support Vector Machines (SVM) is a developed learning system originated from the statistical learning theory [200]. Designing SVM classifiers includes selecting the proper kernel function and choosing the appropriate kernel parameters and C value. The popular library for support vector machines, LIBSVM(`www.csie.ntu.edu.tw/~cjlin/libsvm`) was used in the experiment. We use the Radial Based Function (RBF) kernel for the SVM classifier. The parameter γ that defines the spread of the radial function was set to be 5.0 and parameter C that defines the trade-off between the classifier accuracy and the margin (the generation) to be 3.0.

A random forest (RF) classifier [15] consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. The RF algorithm combines "bagging" idea to construct a collection of decision trees with controlled

variations. In the comparison experiments, the number of trees for random forest classifier was chosen as 300 and the number of variables to be randomly selected from the available set of variables was selected as 20.

As there are only about 20 images in each of the 10 classes of all the image data sets, we designed holdout experiments in the following setting. In each experiment, we randomly picked up 2 samples from each class as a testing and validation, respectively, while leaving the remaining data as training. The classification accuracies are calculated as the averaged accuracies from 100 runs, such that each run used a random splitting of the data.

Fig. 3.11 presents a comparison of the results achieved from each of the above single models on the three microscopy image datasets (RNAi, 2D-Hela and CHO). It appears that for each image dataset, the best result was obtained by using MLP. For RNAi, the best result from MLP is 85.3%, which is better than the published result 82% [113]. The accuracies from other three classifiers are 71.0% (kNN), 72.3% (random forest), and 74.5% (SVM). For 2D-Hela and CHO, the best results obtained by MLP are 84.7% and 93.2%, respectively, which are also very competitive. The results for these two datasets obtained by Shamir et al. are 84% for 2D-Hela and 93% for CHO [113]. The MLP obtained the best performance on the breast cancer biopsy images with classification accuracy of 93.33%. The results obtained by MLP contrast to the generally accepted perception that SVM classifier is better than neural network in classification. The most reasonable explanation for the better performance of MLP from our experiments is that MLP as a memory-based classifier is more resistant to insufficient data amount comparing the margin or distance-based SVM.

In the next experimental part of this study, we seek to show that using random subspace ensemble of MLP can achieve better classification results than the single MLP classifiers used in the previous experiment. And we also try to answer the question that how many MLP should be aggregated in the ensemble to achieve a better result. The result obtained by MLP random subspace ensemble was compared with the results obtained by other two ensemble methods: Dynamic Classifier Selection [207] and Rotation Forest [164].

The settings for all the experiments are as follows: in each run of the experiment, we randomly picked up 80% samples from each class as the training samples, and left 10% samples for validation and 10% for testing, respectively, such that each run used a random splitting of the data. The classification accuracies are calculated as the averaged accuracies from 100 runs. The numbers of MLP tested in the experiment are from 10 to 80. To ensure the diversity among the MLPs in an ensemble, we varied the number of hidden units in the component networks by randomly choosing it from a range of $30 \sim 50$. The classification results obtained by the ensemble has twenty components can be seen in Table 3.2.

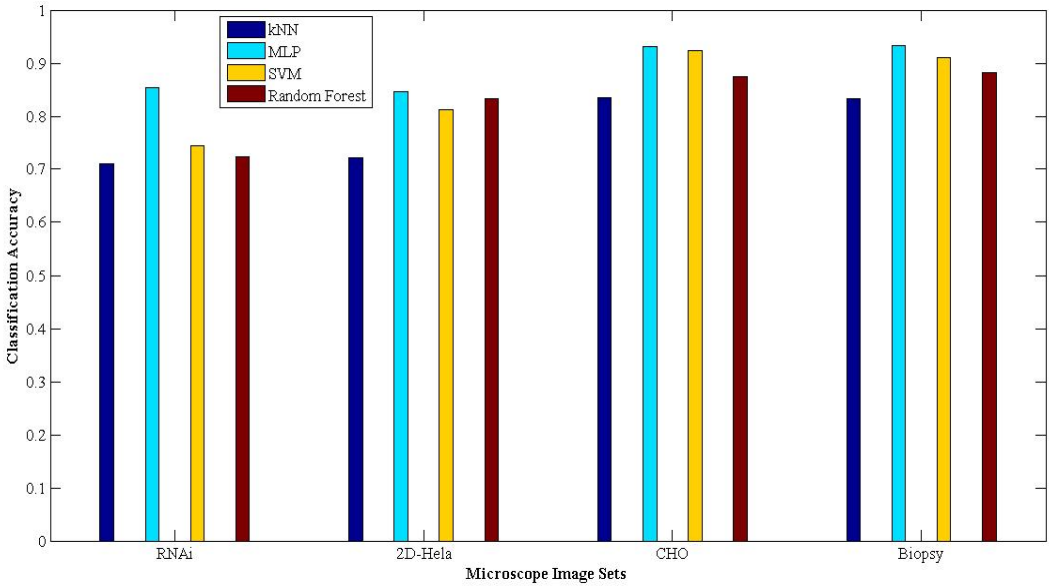


Figure 3.11: Barplots comparing the classification accuracies from four classifiers on microscope image sets

Table 3.2: Improvement of classification accuracy by using Random Subspace MLP Ensemble

Classifier	RNAi	2D-Hela	CHO	Biopsy
MLP	85.30%	84.70%	93.20%	93.33%
MLP-RSE (ensemble size=20)	86.60%	86.30%	93.70%	94.61%

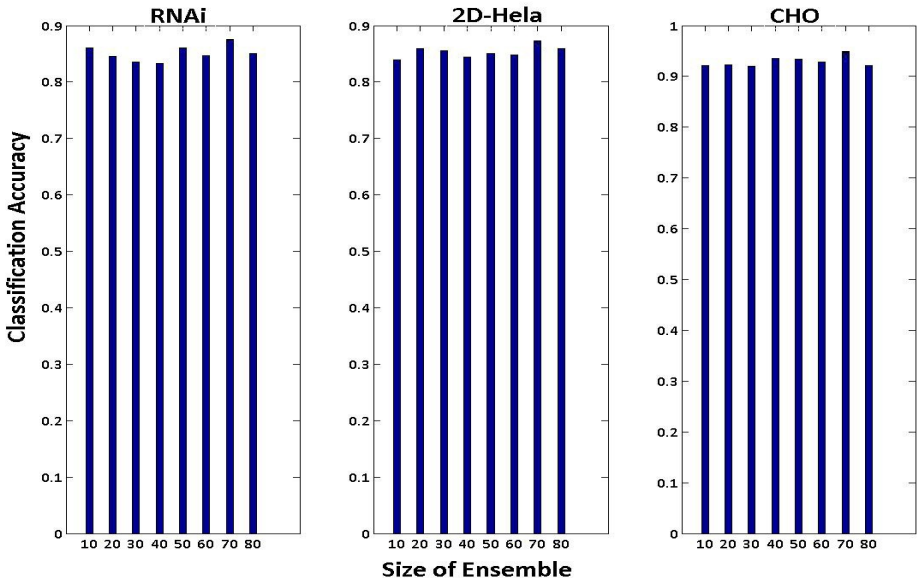


Figure 3.12: Barplots comparing the classification accuracies from different ensemble sizes on fluorescence image sets

From Table 3.2, one can see that for all the four image data sets, the random subspace MLP ensemble does bring the improvement on the classification accuracy, for the RNAi data set, the ensemble brings an increase approaching 1% on classification accuracy, from 85.3% upgraded to 86.6%. The classification accuracies for the other three data sets also be improved, for 2D-hela, it has been enhanced from 84.7% to 86.3%; for CHO, the classification accuracy has been upgraded to 93.7%. The breast cancer biopsy image set achieved 94.61% comparing to the non-ensemble accuracy 93.33%.

To answer the question whether more component neural networks included in an ensemble could further enhance the classification performance, we go on the experiment by varying the sizes of the ensemble from 10 components networks to 80 networks in each of the ensemble. The results of the averaged classification accuracies are shown in Fig. 3.12 and Fig. 3.13. It seems that for fluorescence image sets, bigger ensemble size does bring better classification performance. As can be seen from Fig. 3.12, at the ensemble size 70, for all of the three image data sets, we reach better classification accuracies than other ensemble sizes. But such improvement becomes marginal after the size exceed a limit and the bigger ensemble sizes bring heavy computational burden on the training phase. This is also true for the breast cancer biopsy image set, the best performance for biopsy image set was obtained at the ensemble size 40 (Fig. 3.13. The classification results of these three data sets are enhanced comparing to the results in Shamir et al. 2008 [113].

In Table 3.3, Table 3.4, Table 3.5 and Table 3.6, the top-5 classification results from single MLPs and the best ensemble results were listed for comparison, an apparent

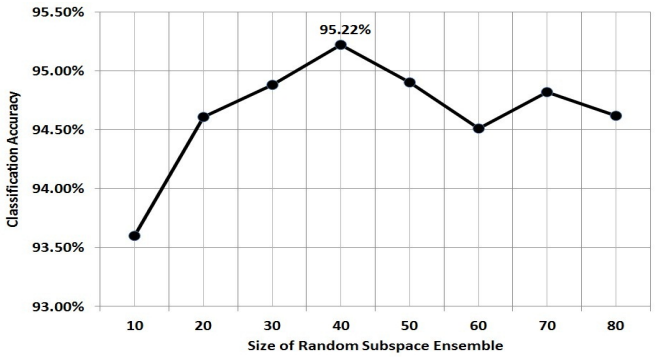


Figure 3.13: Classification accuracies from different ensemble sizes on breast cancer biopsy image set

Table 3.3: Performance from Random Subspace Ensemble of RNAi

Indices	Accuracy (mean)	Standard Deviation
1	86.40%	0.1265
2	86.30%	0.1243
3	85.90%	0.1280
4	85.90%	0.1215
5	85.90%	0.1248
Ensemble (size=70)	87.13%	0.1202

Table 3.4: Performance from Random Subspace Ensemble of 2D-Hela

Indices	Accuracy (mean)	Standard Deviation
1	85.12%	0.0583
2	84.86%	0.0512
3	84.79%	0.0570
4	84.72%	0.0563
5	84.70%	0.0494
Ensemble (size=70)	87.98%	0.0518

Table 3.5: Performance from Random Subspace Ensemble of CHO

Indices	Accuracy (mean)	Standard Deviation
1	93.85%	0.0635
2	93.63%	0.0597
3	93.55%	0.0577
4	93.35%	0.0543
5	93.35%	0.0595
Ensemble (size=70)	94.67%	0.0542

conclusion is that the average classification results of 100 runs obtained by random subspace MLP ensemble are superior than any result obtained by one single MLP, and the ensemble offers relatively smaller standard deviations.

In the following, we evaluated three different types of MLP ensembles for classification. The ensemble methods we compared are Random Subspace, Rotation Forest

Table 3.6: Performance from Random Subspace Ensemble of Breast Cancer Biopsy

Indices	Accuracy (mean)	Standard Deviation
1	94.39%	0.0590
2	94.39%	0.0588
3	94.33%	0.0647
4	94.22%	0.0615
5	94.11%	0.0623
Ensemble (size=40)	95.22%	0.0523

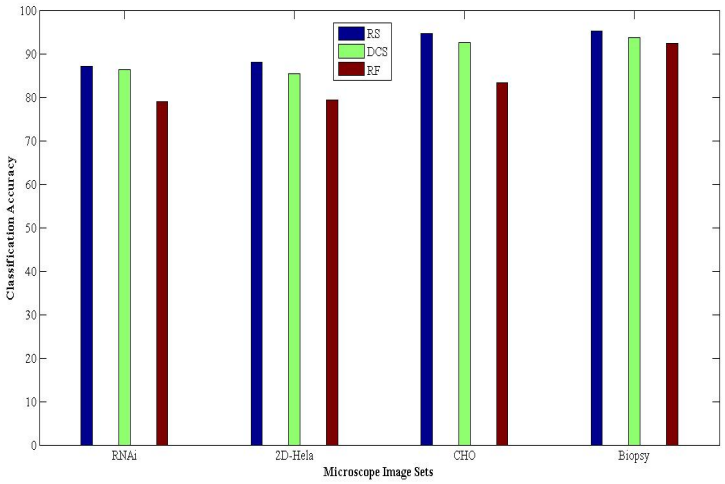


Figure 3.14: Classification accuracies from different ensemble methods on microscope image sets

and Dynamic Classifier Selection. Rotation Forest ensemble and dynamic classifier selection are two ensemble method proposed recently, details of these two methods can be seen in [164] and [207]. The experiment settings for these three ensemble methods are similar, the comparison result of these three ensemble methods are shown in Fig. 3.14. In Fig. 3.14, for RNAi, Hela and CHO image sets, the ensemble size is set as 70, for breast cancer biopsy image set, the comparison is made on size 40.

Although in Fig. 3.14, we only listed the best results under the fixed ensemble sizes, in our experiment we found that for each the ensemble size we tested, the performance of the random subspace ensemble is superior than the performance of rotation forest. In the cases that the ensemble sizes are less than 40, dynamic classifier selection can obtain better result than random subspace, but when the ensemble size keeps growing, random subspace ensemble gave the best classification result among these three methods. The other traditional ensemble methods such as Bagging and Boosting were not included in this comparison since it has been proven that in linear classifier situations, random subspace always give better result than Bagging and Boosting [180].

The confusion matrix that summarizes the details of the above random subspace

Table 3.7: Averaged confusion matrix for RNAi

%	1	2	3	4	5	6	7	8	9	10
1 (CG10873)	0.96	0.02	0	0.02	0	0	0	0	0	0
2 (CG1258)	0.01	0.87	0	0.02	0.03	0	0.07	0	0	0
3 (CG3733)	0	0.01	0.96	0	0	0	0	0.01	0	0.02
4 (CG7922)	0.05	0	0	0.82	0	0	0	0.13	0	0
5 (CG8222)	0	0.03	0	0	0.89	0	0	0	0	0.08
6 (CG12284)	0	0	0	0	0	0.88	0.04	0	0.08	0
7 (CG17161)	0	0.05	0	0	0	0	0.91	0	0.04	0
8 (CG3938)	0.01	0	0	0.12	0.03	0	0	0.84	0	0
9 (CG8114)	0	0.03	0	0.02	0	0.02	0	0	0.93	0
10 (CG9484)	0	0	0.01	0.04	0.11	0.17	0	0.02	0	0.65

Table 3.8: Averaged confusion matrix for 2D-Hela

%	1	2	3	4	5	6	7	8	9	10
1 (Actin)	0.97	0	0	0	0	0	0	0.02	0.01	0
2 (Dna)	0	0.78	0.03	0	0	0.15	0.01	0.03	0	0
3 (Endosome)	0	0.06	0.9	0.02	0	0	0	0.01	0	0.01
4 (Er)	0	0	0	0.85	0.15	0	0	0	0	0
5 (Golgia)	0	0	0	0.14	0.82	0.02	0	0	0.02	0
6 (Golgpp)	0	0.06	0	0	0.04	0.86	0	0.03	0.01	0
7 (Lysosome)	0.01	0.04	0.03	0	0	0.03	0.84	0.04	0	0.01
8 (Microtubules)	0	0.04	0.04	0.03	0	0.03	0.02	0.84	0	0
9 (Mitochondria)	0	0	0	0.02	0	0	0	0	0.98	0
10 (Nucleolus)	0	0	0.02	0.02	0	0	0	0	0	0.96

Table 3.9: Averaged confusion matrix for CHO

%	1	2	3	4	5
1 (Giantin)	0.92	0	0.08	0	0
2 (Hoechst)	0.02	0.98	0	0	0
3 (Lamp2)	0.01	0	0.99	0	0
4 (Nop4)	0	0	0	0.97	0.03
5 (Tubulin)	0.02	0.01	0.03	0.05	0.89

ensemble on RNAi image data set is given in Table 3.7. For the total number of 10 testing samples (one for each category) in each run of the experiment, the 10-by-10 matrix records the number of correct predictions (the diagonal elements in the matrix) and incorrect predictions (the non-diagonal elements) made by the classifier ensemble compared with the actual classifications in the testing data. The matrix are averaged from the results of 100 runs. It is apparent that among the 10 classes, *CG10873*, *CG7922*, *CG1258* and *CG3733* types are the easiest to be correctly classified while the *CG9484* is the difficult category. The confusion matrices for 2D-Hela, CHO and breast cancer biopsy data sets are given in Table 3.8, Table 3.9 and Table 3.10, respectively.

Table 3.10: Averaged confusion matrix for the image dataset (ensemble size=40)

%	Healthy	Tumor insitu	Invasive carcinoma
1 (Healthy)	0.9517	0.0393	0.0090
2 (Tumor insitu)	0.0240	0.9412	0.0348
3 (Invasive carcinoma)	0.0120	0.0243	0.9637

3.7 Conclusion

Ensemble of classifiers is an effective method for machine learning and can improve the classification performance of a standalone classifier. A combination aggregates the results of many classifiers, overcoming the possible local weakness of the individual classifier, thus producing a more robust recognition. In this work, we aimed at improving the challenging multi-class microscopic image classification problem. Two contributions are presented. Firstly, we proposed to apply the combination of curvelet transform, gray level co-occurrence matrix and completed local binary patterns to efficiently describe microscopic images, which exhibit very high directional sensitivity and are highly anisotropic. Secondly, we have examined a novel method to incorporate random subspace based multi-layer perceptron ensemble. The designed paradigm seems to be well-suited to the characteristics of microscopic image data. It has been empirically confirmed that considerable improvement in the classification can be produced by using the random subspace neural network ensembles. Experiments on the benchmark RNAi datasets showed that the random subspace MLP ensemble method achieved higher classification accuracies ($\sim$ 87.1%). Compared to the published result 82%, a 4.9% improvement on the classification accuracy was obtained. The classification results of other three groups of microscopy image data sets using random subspace MLP also support the effectiveness of the proposed method. The random subspace MLP ensemble obtained 86.6% classification accuracy on the 2D Hela dataset, and 93.7% on the CHO dataset, providing the improvements of 0.7% and 2.6% on the classification accuracy, respectively. A classification accuracy of 95.22% was obtained from the proposed ensemble method on the biopsy image sets, which obtains an 1.82% improvement on the published result on the same image sets.

Chapter 4

A Two-stage Classification Scheme for Reliable Breast Cancer Diagnosis

The content of this chapter has been published in the following papers:

- Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu. Breast Cancer Diagnosis from Biopsy Images with Highly Reliable Random Subspace Classifiers Ensemble. Machine Vision and Applications, Vol. 24, No. 7, pp. 1405-1420, 2013.
- Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu. Highly Reliable Breast Cancer Diagnosis with Cascaded Ensemble Classifiers, Proceedings of the International Joint Conference on Neural Networks 2012 (IJCNN 2012), pp. 1-8, Brisbane, Australia, June 2012.

4.1 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 [181]. 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 [3]. 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 [118].

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 ultrasonics. 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 [6]. Among these FNA is the most convenient because it involves the use of very small needles (smaller than those used for blood tests) [16]. 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. The 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 [14, 121, 147]. 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 [187]. The wavelet package transform coupled with local binary patterns were used for meningioma subtype classification in [156]. 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 [68]. Among them, ensemble learning has attracted much attention due to the good performance from many applications in medicine and biology [213]. In the case of ensemble classification, ensemble learning is concerned with the integration of the results of a set of classifiers (often called as 'base classifiers') [108] 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 [78] is often quoted as an efficient way of combining the results of a set of classifiers. A recent application of RS for functional Magnetic Resonance Imaging (fMRI) classification has shown promising results [107]; 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 [11]. 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 achieves the smallest error rate. In many

applications, however, it is more important to address the reliability issue in classifier design by introducing a reject option which allows for an expression of doubt. The objective 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 reject option has been a topic of interest in pattern recognition. Multi-stage classifiers are the ensembles that each individual classifier in the ensemble has a reject option [151]. Cascading [50] is a scheme to support multi-stage classification. Many cascading multi-stage classifier architectures have been proposed [151, 63, 56] and plenty of promising results have been achieved in medical and biological classification applications, such as microarray data classification [141] and gene expression data classification [74].

In this chapter, we propose and evaluate a novel cascade scheme, comprised of two random subspace ensembles, to be applied to microscopic biopsy image classification. 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, 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, an 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.

We also investigated the effectiveness of a feature description approach by combining Local Binary Pattern (LBP) texture analysis, statistics derived from 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.

The rest of this chapter is organized as follows: In Section 4.2, some related works on biopsy image analysis and classification are presented. In Section 4.3, we describe and theoretically analyze the proposed two-stage ensemble cascading system in detail. In Section 4.4, 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 Section 4.5.

4.2 Related Works

The automated classification of biopsy images involves the identification of multiple classes, including benign, cancerous and confounder classes. The energy and entropy features from multiwavelet coefficients of the biopsy images were used in [85], the leave-one-out technique was used for error estimation, a 97% classification rate was reported in their paper. Zhu et al. [230] used fluorescence spectroscopy of breast tissues for 121 biopsy images, the tissue spectra were analyzed using a partial least-squares analysis and a set of Principle Components were used for feature selection. SVM was then used for classification, a cross-validated sensitivity and specificity of up to 81% and 87% was reported. Dalle et al. [41] proposed a multiresolution approach for breast cancer grading. Cells were segmented using Gaussian color models and classified using the Gaussian distribution. Doyle et al. [45] used a combination of graph-based, morphological and textural features for prostate cancer classification. The SVM classifier was used and an accuracy of 92.8% reported when distinguishing between Gleason grade 3 and Stroma. The aggregation of color, texture and morphology features were also used by Tabesh et al. [187] for prostate cancer biopsy image classification, the mixed features together with linear Gaussian and kNN classifiers achieved an accuracy of 96.7%. Basavanhally et al. [9] investigated lymphocytic infiltration in HER2+ breast cancer, a total of 50 image-derived features describing the arrangement of the lymphocytes were extracted from each biopsy image, a classification accuracy of 90% was obtained by SVM. In [49], a Multiple Instances Learning SVM (MILSVM) was proposed for intraductal breast lesion classification, quantitative features of 327 regions of interests from 62 patient biopsy cases was used for classifier training, 84.6% classification accuracy was obtained from 149 test ROIs. A cascade classification scheme for prostate cancer biopsy images was proposed in [48], the biopsy cases were first classified into cancerous and non-cancerous cases, then a grading system was used to categorize the cancerous cases into different cancer grades, a positive predictive value of 86% was reported. Krishnan et al. [103] extracted textural features of images to train and select the best classifier from five different kinds of classifiers, the best recorded classification accuracy was

95.7% obtained from the combined features coupled with fuzzy classification. In most of these papers, the authors provide a consensus that using multiple image features is an effective way for biopsy image classification. In a more recent study by Kothari et al. [101], Fourier shape descriptors were used to capture the distribution of stain-enhanced cellular and tissue structures, the authors claimed that the Fourier shape descriptors produced better performance than other textural image features, however they also admitted that the time cost for their algorithm was much higher than in the case of other feature extractors.

Due to the multiple image scales at which relevant information may be extracted from biopsy images, the use of an ensemble of classifiers as opposed to an individual classifier has been proposed. A multiclass system was used by Sboner et al. [172] for skin biopsy image classification, 38 geometric and colorimetric features were extracted from digital images of skin lesions, three different kinds of classifiers, namely linear discriminant analysis (LDA), k-NN and a decision tree classifier were combined to produce a final classification result using a voting scheme. This work suggested that a suitable combination of different kinds of classifiers can improve the performance of an automatic diagnostic system. A local patch-based subspace ensemble method was proposed in [120] for brain MRI image classification, which built multiple individual classifiers, based on different subsets of local patches, and then combined them for more accurate and robust classification. They obtained a 90.8% classification accuracy, demonstrating a very promising performance compared with other state-of-the-art methods for AD/MCI classification of MR images. Doyle et al. [46] presented a boosted Bayesian multiresolution (BBMR) system to identify regions of prostate cancer on digital biopsy slides. The Adaboost ensemble method was used for feature selection. Their experimental results demonstrated that the proposed system outperformed individual classifiers and a Bagging Random Forest.

4.3 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 have 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 an individual patient. In such cases an overall quality measurement of a classifier (e.g. classification accuracy) would not provide the desired information, even where good accuracies are achieved using some state-of-art methods. 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 chapter a new two-stage classification method for biopsy image classification, consisting of a random subspace ensembles with reject option, is proposed. We adopted the definitions of recognition rate, rejection rate and reliability proposed in [220], 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- no. of rejected 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.

According to this definition of reliability, high reliability can be achieved with an appropriate trade-off between error rate and rejection rate.

4.3.1 Reject Option for Classification

The optimal classification rule with reject option was defined by Chow [30]. 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 should be 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:

$$f(x) = \begin{cases} argmax_{C_i}(p(C_i|x)) & \text{if } max_{C_i}\ (p(C_i|x)) \geq t \\ \\ reject & \text{if } \forall_i\ p(C_i|x) < t \end{cases} \tag{4.1}$$

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

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

$$p(reject) = \int_{reject} p(x)dx = p(max(p(C_i|x)) < t). \tag{4.2}$$

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.

4.3.2 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. 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 chapter, we will propose a hybrid classification scheme which serially connects two parallel random subspace ensembles of classifiers (Fig. 4.1). Note that all classifiers have a reject option.

In our current implementation the first ensemble (Classifier Ensemble 1 in Fig. 4.1) consists of a collection of SVM classifiers, the second (Classifier Ensemble 2 in Fig. 4.1) consists of a collection of MLP classifiers. From Fig. 4.1 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".

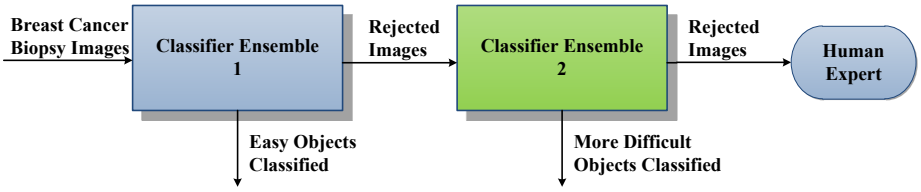


Figure 4.1: Operation of the hybrid classification scheme comprising a cascade of two Random Subspace classifier ensembles.

SVM and MLP have obtained satisfactory performance in many medical image analysis tasks, especially in histopathological image analysis [69], 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 [200]. 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 [33]. 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 [108]. Making use of different individual classifiers in an ensemble can improve the performance, here we expand the concept to employ different base classifiers for the two ensembles to enhance the "diversity" between the ensembles.

The major issue for designing the above 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 [30] and [193]. Intuitively, a suggested classification should be rejected if the confidence in that classification is below a 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. 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-versus-all* approach where each classifier classifies records as belonging or not belonging to a class). The desired multiclass 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 [189]:

$$P(y = +1|\mathbf{x}) = \frac{1}{1 + \exp(\mathrm{a}\rho(\mathbf{x}) + \mathrm{b})} \tag{4.3}$$

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 $\rho(x)$ [57].

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

$$Class = \arg\max \ P_i, \quad i = 1, \ldots, K \tag{4.4}$$

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 [78]. 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 uses each one to train 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 SVM 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.

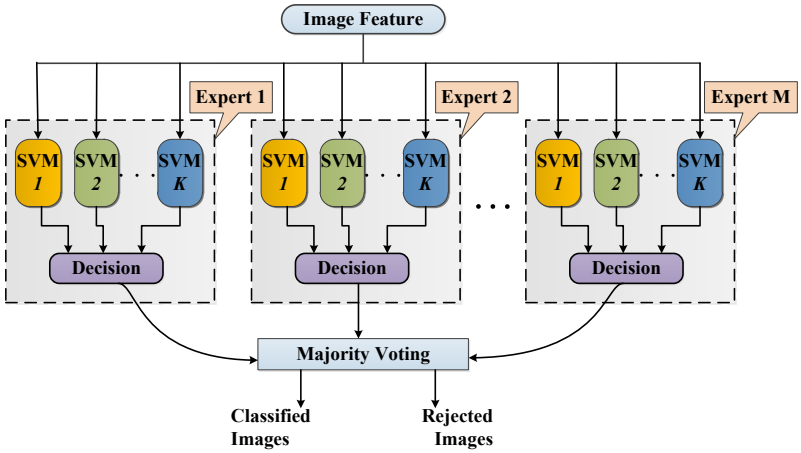


Figure 4.2: SVM ensemble with rejection option in stage 1, which consists of a set of binary SVMs (experts)

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 [114]. Fig. 4.2 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 K 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. 4.3. In combining the decisions from the M base MLPs, a sample (selected from the collection of 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.

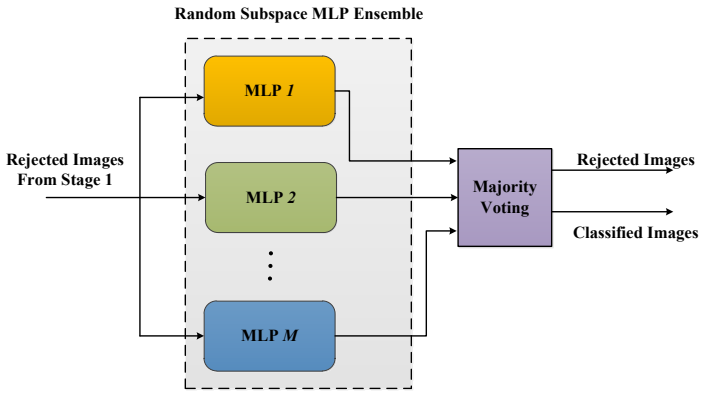


Figure 4.3: Illustration of the stage 2 Random Subspace classifier ensemble which consists of a set of MLPs

4.3.3 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:

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)} = \frac{p(x|C_i)p(C_i)}{\sum_{i=1}^{M} p(x|C_i)p(C_i)} \tag{4.5}$$

where M 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 [63] 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. Assume $MAX_p^1 = max_{j\in[1,\ldots,N]}p(C_j|x)$ 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:

$$\epsilon_1 = \int_A (1 - MAX_S^1)p(x)dx \tag{4.6}$$

where A is the region composed of all accepted instances. Using some simple manipulations on Equation 4.6 , we then get the following:

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

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 $\int(1 - p(x))p(x)dx$. The second term comes from the errors generated during stage 1. This situation can be illustrated as in Fig. 4.4, 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.

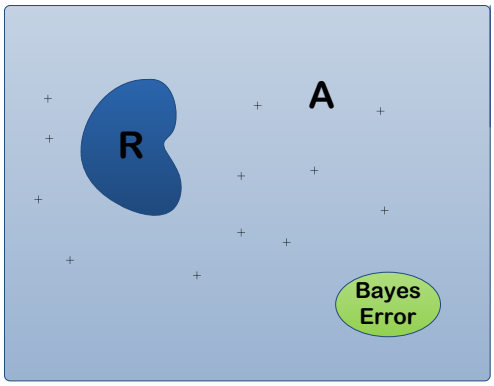


Figure 4.4: Error rate of stage 1

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:

$$\begin{aligned}\epsilon_2 &= \int_R (1 - MAX_p^2)p(x)dx \\ &\quad + \int_{R\cap I^M} (MAX_p^2 - MAX_S^2)p(x)dx \end{aligned} \tag{4.7}$$

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:

$$\begin{aligned}\epsilon &= \epsilon_1 + \epsilon_2 \\ &= \int_A (1 - MAX_p^1)p(x)dx + \int_R (1 - MAX_p^2)p(x)dx \\ &\quad + \int_{A\cap I^S} (MAX_p^1 - MAX_S^1)p(x)dx \\ &\quad + \int_{R\cap I^S} (MAX_p^2 - MAX_S^2)p(x)dx \\ &= \epsilon_{Bayes} + \int_{A\cap I^S} (MAX_p^1 - MAX_S^1)p(x)dx \\ &\quad + \int_{R\cap I^M} (MAX_p^2 - MAX_S^2)p(x)dx. \end{aligned} \tag{4.8}$$

From Eqn. 4.8, for approaching the goal that $\epsilon = \epsilon_{Bayes}$, 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 [220], 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.

4.4 Experiments

MATLAB 7.0 was used to implement the algorithms 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.

4.4.1 Image Sets and Feature Extraction

One breast cancer benchmark biopsy image 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. For the details of the image sets, see Section 3.3.2.

Three image feature extractors were used for quantitatively describing biopsy images. Local Binary Patterns (LBPs) was used for extracting local textural features, Gray Level Co-occurrence Matrix (GLCM) statistics for representing global textures and the Curvelet Transform for multiresolution shape description. See Section 3.4 for details of these feature extraction methods.

The mean μ, the standard deviation δ and the entropy H for each curvelet sub-band are used as the curvelet features. If n curvelets are used for the transform, $3n$ features $G = [G_\mu, G_\delta, H]$ are obtained, where $G_\mu = [\mu_1, \mu_2, \cdots, \mu_n]$, $G_\delta = [\delta_1, \delta_2, \cdots, \delta_n]$ and $H = [h_1, h_2, \cdots, h_n]$. A $3n$ dimensional feature vector can be used to represent each image in the dataset. 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 3.1 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.

4.4.2 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 Section 2. The six classifiers were (i) kNN, $k = 3$, (ii) single MLP, (iii) single SVM, (iv) Logistic Regression, (v) Fisher Linear Discrimination and (vi) Naive Bayes Classifier [165]. 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

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

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. 4.5, 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. 4.5. 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.

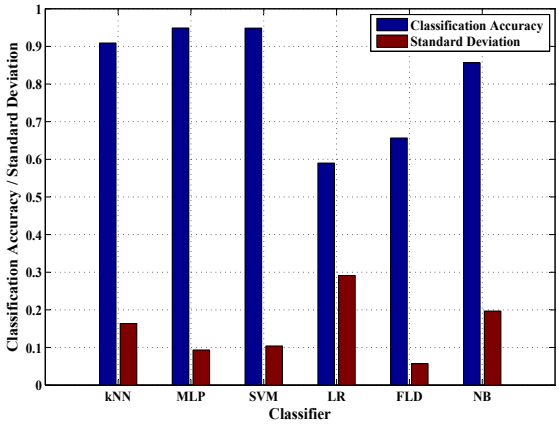


Figure 4.5: Classification accuracies and standard deviations from applying kNN, single MLP, single SVM, Logistic Regression (LR), Fisher Linear Discrimination (FDL), and Naive Bayesian (NB)

Fig. 4.6 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 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

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

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 Bayes 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. The experimental results are consistent with other research works, that in general SVM and MLP can achieve better classification performance on biopsy image analysis [69].

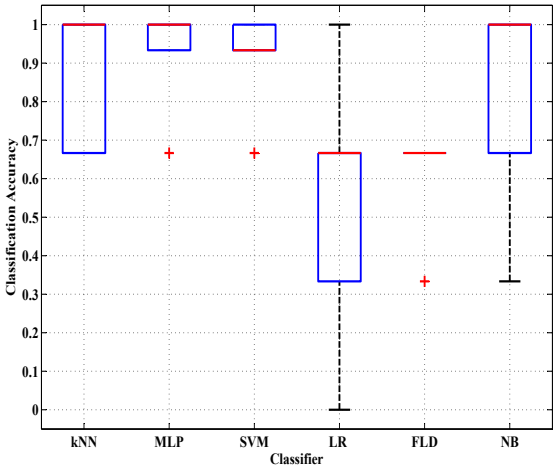


Figure 4.6: Boxplot of classification accuracies from applying single MLP, single SVM expert, Random Subspace SVM ensemble (RS-SVM) and Random Subspace MLP ensemble (RS-MLP)

4.4.3 Evaluation of Random Subspace Ensembles

Table 4.1 shows the classification accuracies obtained using 7 different ensemble classifiers with different image feature combinations. The classifier ensemble methods compared here are: (i) Bagging [111] with SVM (BagSVM), (ii) Bagging with MLP (BagMLP), (iii) AdaBoost [211] with SVM (BoostSVM), (iv) AdaBoost with MLP (BoostMLP), (v) Random Forest [15] with decision trees (RandF), (vi) Random Subspace with MLP (RSMLP) and (vii) Random Subspace with SVM (RSSVM). The three different image feature types introduced earlier were considered: Curvelet, GLCM, and LBP, which are represented by the letters C, G, and L in Table 4.1 respectively. Each image has a 666 dimensional feature vector with all of these three features. Each randomly selected subspace used 80 percent 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 4.1). 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 4.1.

Table 4.1: 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

One can note from Table 4.1 that the use of ensembles does improve the classification accuracy. RSSVM and RSMLP produced the best performance regardless of the types of image features used for the training, both obtained classification accuracies around 95% with the combined feature (C&G&L), 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 the same image dataset from using other kinds of features are also compared in the experiment, as in [16], the level set method was used to extract image features, and a 42-bins 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 [107], 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 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 Fig. 4.7 and Fig. 4.8, respectively.

The same conclusion as in [63] can be drawn from Fig. 4.7 and Fig. 4.8. 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 Sub-

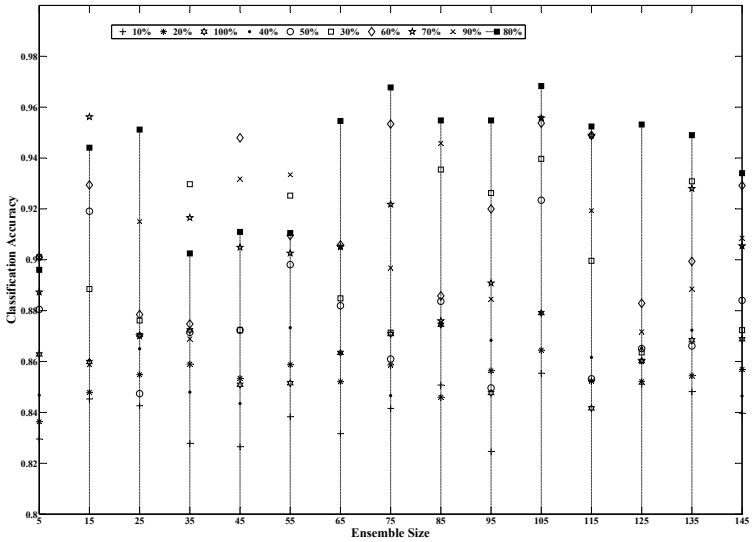


Figure 4.7: Classification results of the RSSVM ensemble with different ensemble sizes and different cardinalities of training feature

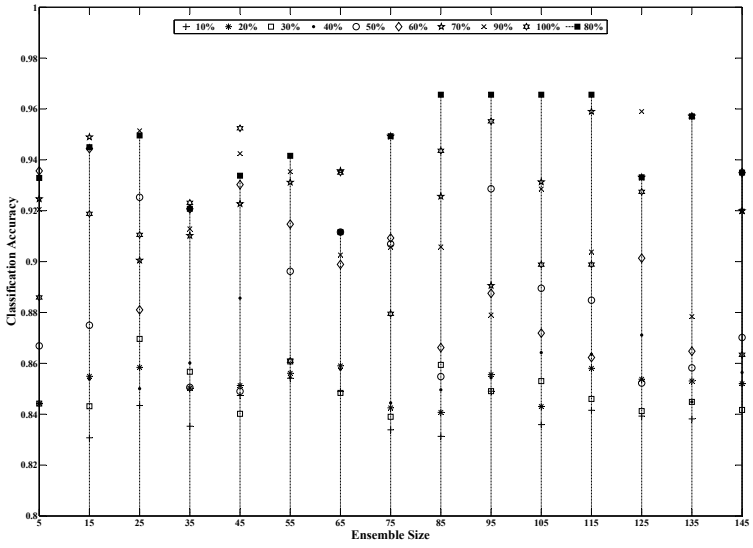


Figure 4.8: Classification results of the RSMLP ensemble with different ensemble sizes and different cardinalities of training feature

space 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.

4.4.4 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 determined as 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 (stage 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 ensure 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 4.2.

Table 4.2: Classification Accuracy and Reliability of Different Cascade Schemes on the Biopsy Image Data with rejection threshold of both stages equal to 84, RR stands for Recognition Rate, Re for Reliability, ReR for Rejection Rate, and ER represents Error Rate, see Section 3 for details

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

From Table 4.2, 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 [63].

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 Fig. 4.9 and Fig. 4.10, 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.

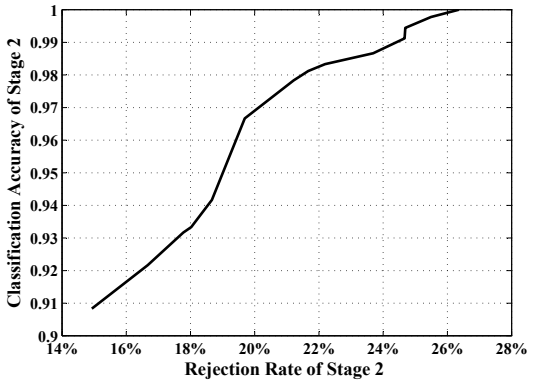


Figure 4.9: Averaged stage 2 accuracies with 10 varying stage 2 rejection rates

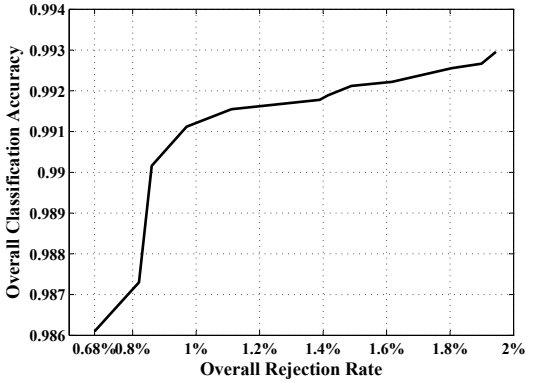


Figure 4.10: Averaged overall classification performances from 10 varying overall rejection rates

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 4.3. Compared with the results in Table 4.2, 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.

Table 4.3: 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

The confusion matrix from the overall performance that summarize the detailed situations of rejection rate 1.94% were displayed in the Table 4.4. 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.

4.4.5 Results on UCI Datasets

In order to further evaluate our proposed system, we compared our proposed method with Negative Correlation Learning (NCL) proposed in [17], which is also a neural network ensemble classifier, the classifiers in the ensemble are trained with NCL. For the two methods compared here, we fixed the ensemble sizes as 105. The rejection threshold for stage 1 and stage 2 were set as $t_1 = 84$ and $t_2 = 95$ for our two ensembles trained with Random Subspace.

Table 4.5 shows the classification error rates of two empirical tests, on the Wisconsin breast cancer dataset from the UCI repository (699 patterns), and the Heart disease dataset from Statlog (270 patterns).

Table 4.4: 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

Table 4.5: Averaged Error Rate of Two Methods on Two UCI Datasets (%)

Dataset	NCL	Proposed
Breast Cancer	3.12	**0.74** (with rejection rate 0.89%)
Heart Disease	17.33	**14.54** (with rejection rate 1.67%)

4.5 Conclusion and Future Work

In this chapter, 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%). The proposed method obtained a 5.6% improvement on the classification accuracy compared with the best published result. 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 issues 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 chapter, 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.

Chapter 5

Cascading One-Class Kernel Subspace Ensembles for Reliable Medical Image Classification

The content of this chapter has been published in the following papers:

- Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu. One-Class Kernel Subspace Classifier Ensemble for Medical Image Classification, Eurasip Journal on Advances in Signal Processing, 2014:17, pp. 1-13, 2014.
- Yungang Zhang, Bailing Zhang, Frans Coenen and Wenjin Lu. Cascading One-Class Kernel Subspace Ensembles for Reliable Biopsy Image Classification, Journal of Medical Imaging and Health Informatics, Vol.4, pp. 1-12, 2014.

5.1 introduction

In many automatic medical diagnosis applications, the datasets used for diagnosis is often imbalanced as the number of normal cases is usually larger than the number of the disease cases. Classifiers that generalize well over balanced data are not the most appropriate choice in such an unbalanced situation. For example, decision trees tend to over-generalize the class with the most examples; Naive Bayes requires enough data for the estimation of the class-conditional probabilities [119]. One-Class Classifiers (OCC) [192] are more appropriate for such a task.

Using of a single classifier often fails to capture all aspects of the data in many real classification tasks, therefore, a combination of classifiers (an ensemble) is often considered to be an appropriate mechanism to address this shortcoming. The main idea behind the ensemble methodology is to use several classifiers, and combine the individual results in order to produce a classification that outperforms the outcomes that would have been produced were the classifiers to operate in isolation [166]. Ensembles of one-class classifiers have also been shown to perform better than when using individual classifiers [65, 10, 72]. There are many strategies for constructing a classifier ensemble,

examples include: using different training data sets, different feature subsets, various types of individual classifiers and different fusion rules. Among these, the feature subset strategy has shown better performance when the dimensionality of the feature vector is high compared to the number of the data samples [157, 104, 219, 215]. It is thus suggested that the feature subset ensemble strategy is consequently well suited to medical image classification problems, as various types of image features are generally extracted for medical image classification tasks, which in turn means that the dimensionality of the vector space is typically larger than the number of image samples, i.e., the "curse of dimensionality". Using the feature subset strategy can avoid such a problem.

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 [151]. Cascading [50] is a scheme to support multi-stage classification. At the first stage of a cascading system, a generalized classifier is used, for each pattern, a classification confidence is given by the system, the patterns with low confidence will not be classified, instead, the system will pass on these uncovered patterns to the next stage. At the next stage, a more complex rule is constructed to focus on these uncovered patterns.

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 classification reliability issue by introducing a reject option which provides for an expression of doubt. The objective 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.

In this chapter, we propose and evaluate a novel classification scheme for breast cancer biopsy images. To stress the reliability of the automatic medical diagnosis, the proposed classification scheme utilizes a cascade of two classifier ensembles. The first stage of the cascade consists of an ensemble of One-Class Classifiers, the ensemble is built with the feature subset strategy; each One-Class classifier is trained with one type of features extracted from the biopsy image training set. The Kernel Principle Component Analysis (KPCA) model was chosen as the base classifier of the first stage. For each image category, n KPCA models can be trained from n types of image features. Therefore, the ensemble size of the first stage is determined by both the number of image classes and the number of image feature types, for example, given a m-class classification task and n different kinds of image features, then the ensemble will consist of $m \times n$ KPCA models. Given an unlabeled image, its n types of features will first be mapped into the kernel space by the corresponding n trained KPCA models from each class.

The mapped features will then be reconstructed from the high dimensional kernel space into the original space by Preimage learning [110], the distances between the original features and the reconstructed features will be measured. The distances given by the KPCA models will be combined to output a confidence score describing the probability of the sample belonging to a class. For a m-class classification task, m confidence scores will be obtained, one for each class. Then a rejection rule will be used to judge if the image should be classified or rejected and passes on to the next stage for further consideration.

The second stage consists of a random subspace [78] ensemble of Support Vector Machines (SVM) which operate using majority voting, any samples that have a low consensus degree will be rejected for further consideration by human experts. The classification with the proposed cascaded ensembles will provide an efficient means to simultaneously reduce the error rate and enhance the reliability by controlling the reliability-rejection trade-off. The proposed classification system was evaluated on two medical image datasets, promising results were obtained.

The rest of this paper is organized as follows: Some related work is considered in Section 5.2. In Section 5.3, we described the proposed two-stage ensemble cascading system in detail. In Section 5.4, some experimental results are presented based on two synthetic datasets and the adopted two real medical image datasets. The paper ends with some conclusions in Section 5.5.

5.2 Related Works

In this section, we will first introduce some related works on one-class classification. Then one-class classifier ensembles will be discussed.

5.2.1 One-Class Classification

The term of One-Class Classification was first proposed by Moya et al. [137]. Many approaches to one-class classification have been presented in the literature [192]. Following the taxonomy in the survey papers of [96, 130, 131], the algorithms used in OCC can be categorized as follows: (i) boundary methods, (ii) density estimation and (iii) reconstruction methods.

Tax and Duin tried to separate the positive class form all other patterns in the pattern space; the positive class data was surrounded by a hyper-sphere which encompassed almost all positive patterns within the minimum radius [189, 191]. The proposed Support Vector Data Description (SVDD) tries to separate the pattern space with data from the space containing no data. Manevitz and Yousef [129] proposed another version of one-class SVM to identify the outlier data as representative of the second class,

and applied their method to the standard *Reuters*[1] dataset and noted that their SVM methods was quite sensitive to the choice of representation and kernel. Although One-class classifiers, such as OCSVM, have been widely used, the estimated boundary can be sensitive to the nature of the data [169]. When noisy data, or many outliers, are contained in the training set, OCSVM will estimate a large boundary that encloses regions of the feature space where the positive class has low density, often resulting in many false positives [79]. This can be highly problematic for many applications, especially for medical diagnosis where the number of false positives must be kept to a minimum, since an accidental diagnosis of a patient as healthy may result in serious consequences.

Density estimation methods estimate the density of the target class to form a model with which to represent the data. Density estimation methods work well if the number of training samples is sufficient enough to estimate data distributions. However, when the models cannot fit the data distribution very well, a large bias may be generated. Details and some comparisons of these methods can be found in [162, 202].

When it is not feasible to obtain large training sets, the reconstruction models can be used to approximate the target class. The reconstruction models aim to produce prototypes of the original data, new objects are projected onto the prototypes. The distance between the original object x and the projected object $p(x)$ (Reconstruction Error), indicates the similarity of a new object to the original target distribution. When the training data has a very high dimensionality, some distance based methods like nearest neighbor tend to perform badly [12]. In such cases it can often be assumed that the target data is distributed in subspaces of much lower dimensionality. Principle Component Analysis [186] is a linear model that has the ability to project the original data into orthogonal space which can capture the variance in the data. In order to approximate nonlinear data distributions, many nonlinear subspace models have also been proposed, such as Self-Organizing Map (SOM), auto-encoders, auto-associative networks and Kernel PCA.

5.2.2 Ensemble of One-Class Classifiers

The existing classifier combination strategies can also be used in one-class classifiers. However, since there is only information from one class, it is more difficult to combine one-class classifiers. Tax and Duin investigated the influence of feature sets and the types of one-class classifiers for the best choice of the combination rule [190]. A bagging based one-class support vector machine ensemble method was proposed in [178]. A dynamic ensemble strategy based on Structural Risk Minimization [86] was proposed by Goh et al. for multiclass image annotation [65]. Recently, some research results have revealed that creating a one-class classifier ensemble from different feature subsets

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578

can provide better performance. Perdisci et al. [152] also used an ensemble of one-class SVMs to create a "high speed payload-based" anomaly detection system, the features were first extracted and clustered, the OCSVM ensemble was then constructed based on the clustered feature subsets. A biometric classification system combining different biometric features was proposed by Bergamini et al. [10], where the one-class SVMs in the ensemble were trained by the data from different people. The feature subset strategy provides diversity with respect to the base classifiers.

Combining one-class classifiers has also shown promising performance in medicine and biology [213]. Peng Li et al. [116] proposed a multi-size patch-based classifier ensemble, which provides a multiple-level representation of image content, the proposed method was evaluated on colonoscopy images and ECG beat detection [115]. The k-nearest neighbor classifier was selected as the base classifier in the work of Okun and Priisalu [146]; majority voting was chosen as the combination rules for the ensemble; the method was evaluated on gene expression cancer data.

5.3 Serial Fusion of One-Class Kernel Subspace Ensembles

Although many supervised learning algorithms, such as neural networks and SVM, have been extensively applied to many medical image classification problems, few of them have 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 an 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 methods. 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 chapter a new two-stage classifier for medival image classification is proposed. In the first stage, an ensemble of Kernel PCA models are combined to determine if an image should be classified or rejected, the KPCA models are trained individually from different image features. The rejected images will further be investigated in the second stage, which is an ensemble of '*one-versus-all*' Support Vector Machines, based on the rejection option, the images will either be classified at this stage or deferred for classification by a human expert (Fig. 5.1).

The construction of the KPCA ensemble will be first introduced in Section 5.3.1, then the reject option for classification will be discussed. The SVM ensemble for the second stage will be considered in Subsection 5.3.2 below.

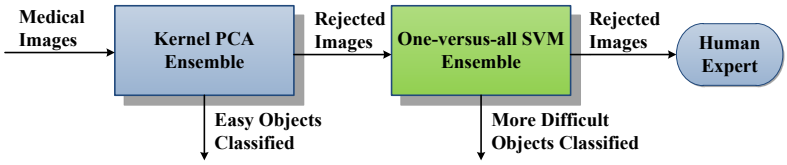


Figure 5.1: Operation of the proposed hybrid classification scheme comprised of a cascade of two classifier ensembles.

5.3.1 One-Class Kernel PCA model Ensemble

In this section the one-class kernel PCA model ensemble will be introduced. This ensemble is the first stage of the proposed cascaded classification system. An individual KPCA model in the ensemble is trained based on an individual image feature. When a new image is to be classified, its features will be reconstructed by corresponding KPCA models, the reconstruction errors from all KPCA models will then be combined and a rejection option will be used to determine whether the image should be classified or rejected.

The theory of Kernel PCA and pattern reconstruction via pre-image will first be introduced, then the proposed KPCA ensemble will be described.

KPCA and Pattern Reconstruction via Pre-image

The traditional (linear) PCA tries to preserve the greatest variations of data by approximating data in a principle component subspace spanned by the leading eigenvectors, noises or less important data variations will be removed. Kernel PCA inherits this scheme, however kernel PCA performs linear PCA in the kernel feature space $\mathbb{H}_\kappa$. Suppose $\mathbb{X} \subset \Re^n$ is the original input data space, $\mathbb{H}_\kappa$ is a Reproducing Kernel Hilbert Space (RKHS) (also called feature space) associated to a kernel function $\kappa(x, y) =< \varphi(x), \varphi(y) >$, where $x, y \in \mathbb{X}$. $\varphi(\cdot)$ is a mapping induced by κ that $\varphi(x) : \mathbb{X} \to \mathbb{H}_\kappa$. Given a set of patterns $\{x_1, x_2, \ldots, x_N\} \in \mathbb{X}$. Kernel PCA performs the traditional linear PCA in $\mathbb{H}_\kappa$. The same as the linear PCA, KPCA also has the eigen decomposition:

$$HKH = U\Lambda U' \tag{5.1}$$

where K is the kernel matrix such that $K_{ij} = \kappa(x_i, x_j)$, and

$$H = I - \frac{1}{N}\mathbf{1}\mathbf{1}' \tag{5.2}$$

is the centering matrix, where I is the $N \times N$ identity matrix, $\mathbf{1} = [1, 1, \ldots, 1]'$ is an $N \times 1$ vector, $U = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N]$ is the matrix containing eigenvectors $\boldsymbol{\alpha}_i = [\alpha_{i1}, \ldots, \alpha_{iN}]'$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_N)$ contains the corresponding eigenvalues.

Denote the mean of the φ-mapped patterns by $\bar{\varphi} = \frac{1}{N}\sum_{j=1}^{N}\varphi(x_j)$. Then for a mapped pattern $\varphi(x_i)$, the centered map $\tilde{\varphi}(x_i)$ can be defined as:

$$\tilde{\varphi}(x_i) = \varphi(x_i) - \bar{\varphi}. \tag{5.3}$$

The kth eigenvector V_k of the covariance matrix in the feature space is a linear combination of $\tilde{\varphi}(x_i)$:

$$V_k = \sum_{i=1}^{N} \alpha_{ki}\tilde{\varphi}(x_i) = \tilde{\boldsymbol{\varphi}}\boldsymbol{\alpha}_k, \tag{5.4}$$

where $\tilde{\boldsymbol{\varphi}} = [\tilde{\varphi}(x_1), \tilde{\varphi}(x_2), ..., \tilde{\varphi}(x_N)]$. If we use β_k to denote the projection of the φ-image of a pattern x onto the kth component V_k, then:

$$\begin{aligned} \beta_k &= \tilde{\varphi}(x)'V_k = \sum_{i=1}^{N} \alpha_{ki}\tilde{\varphi}(x)'\tilde{\varphi}(x_i) \\ &= \sum_{i=1}^{N} \alpha_{ki}\tilde{\kappa}(x, x_i), \end{aligned} \tag{5.5}$$

where:

$$\begin{aligned} \tilde{\kappa}(x,y) &= \tilde{\varphi}(x)'\tilde{\varphi}(y) \\ &= (\varphi(x) - \bar{\varphi})'(\varphi(y) - \bar{\varphi}) \\ &= \kappa(x,y) - \frac{1}{N}\mathbf{1}'\mathbf{k}_x - \frac{1}{N}\mathbf{1}'\mathbf{k}_y + \frac{1}{N^2}\mathbf{1}'\mathbf{K}\mathbf{1} \end{aligned} \tag{5.6}$$

where $\mathbf{k}_x = [\kappa(x, x_1), \ldots, \kappa(x, x_N)]'$. Denote

$$\begin{aligned} \tilde{\kappa}_x &= [\tilde{\kappa}(x, x_1), \ldots, \tilde{\kappa}(x, x_N)]' \\ &= \mathbf{k}_x - \frac{1}{N}\mathbf{1}\mathbf{1}'\mathbf{k}_x - \frac{1}{N}\mathbf{K}\mathbf{1} + \frac{1}{N^2}\mathbf{1}\mathbf{1}'\mathbf{K}\mathbf{1} \\ &= \mathbf{H}(\mathbf{k}_x - \frac{1}{N}\mathbf{K}\mathbf{1}), \end{aligned} \tag{5.7}$$

then β_k in Eqn.(5.5) can be rewritten as: $\beta_k = \boldsymbol{\alpha}_k^{'}\tilde{\kappa}_x$.

Therefore, the projection $P(\varphi(x))$ of $\varphi(x)$ onto the subspace spanned by the first M eigenvectors can be obtained by:

$$\begin{aligned} P(\varphi(x)) &= \sum_{k=1}^{M} \beta_k V_k + \bar{\varphi} = \sum_{k=1}^{M} (\boldsymbol{\alpha}_k'\tilde{\kappa}_x)(\tilde{\boldsymbol{\varphi}}\boldsymbol{\alpha}_k) + \bar{\varphi} \\ &= \tilde{\boldsymbol{\varphi}}\mathbf{L}\tilde{\kappa}_x + \bar{\varphi}, \end{aligned} \tag{5.8}$$

where $\mathbf{L} = \sum_{k=1}^{M} \boldsymbol{\alpha}_k\boldsymbol{\alpha}_k'$.

PCA is a simple method whereby a model for the distribution of training data can be generated. For linear distributions, PCA can be used, however many real world problems are nonlinear. Methods like Gaussian Mixture Models and auto-associative neural networks have been used for nonlinear problems. These methods, however, need to solve a nonlinear optimization problem and are thus prone to local minima and sensitive to the initialization [79]. KPCA runs PCA in the high dimensional feature space through the nonlinearity of the kernel, this allows for a refinement in the description of the patterns of interest. Therefore, Kernel PCA was chosen to model the non-linear distribution of the training samples here.

Kernel PCA has been widely used for classification tasks. A straightforward method using Kernel PCA for classification is to directly use the distances between the mapped patterns in the feature space $\mathbb{H}_\kappa$ to obtain the classification boundaries [174, 79]. However as pointed out in [79], for Kernel PCA, their experimental results showed that the classification performance highly depends on the parameters selected for the kernel function, and there is no guideline for parameter selection in real classification tasks. It is also demonstrated in a more recent work that it is not sufficient to use feature space distance for unsupervised learning algorithms, the distances in the input space are more appropriate for classification [94].

In this paper, we focus on the distances between a pattern x and its reconstruction results by the kernel PCA models trained from different classes. As kernel PCA is used as an one-class classifier here, which means for each class, at least one KPCA model is trained. Suppose there is an m-class classification task, there will be m KPCA models, one for each class. Given an unlabeled pattern x, every KPCA model will produce a projection $P(\varphi(x))_i$, $i = 1, \ldots, m$. During classification, x will be reconstructed in the input space by every $P(\varphi(x))_i$, then m reconstruction results $x_1^{'}, \ldots, x_m^{'}$ can be obtained, the distance between x and each $x_i^{'}$ (also called reconstruction error) is calculated, x will be assigned to the class whose KPCA model produces the minimum reconstruction error. Ideally, the KPCA model trained from the class which x also belongs to will always give the minimum reconstruction error. In our proposed classification scheme, multiple KPCA models are trained for each class, the reconstruction errors of KPCA models from different classes are combined for classification.

In order to obtain the input-space distance between x and its reconstruction result, it is necessary to map $P(\varphi(x))$ back into the input space. The reverse mapping from feature space back to input space is called the *preimage* problem (Fig. 5.2). However, the preimage problem is ill-posed, the exact preimage $x^{'}$ of $P(\varphi(x))$ in the input space does not exist [134], instead, one can only find an approximation $\hat{x}$ in the input space such that

$$\varphi(\hat{x}) = P(\varphi(x)). \tag{5.9}$$

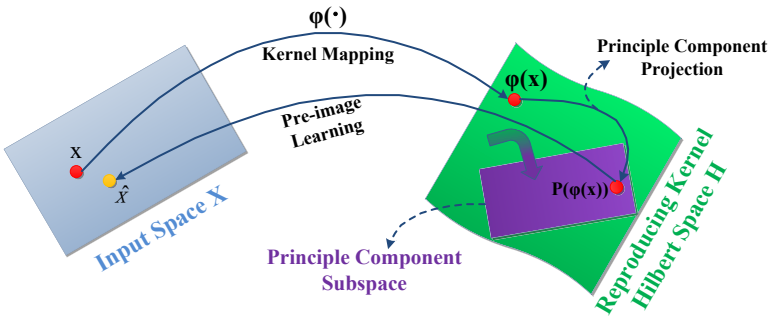


Figure 5.2: Illustration of KPCA preimage learning: the sample x in the original space is first mapped into the kernel space by kernel mapping $\varphi(\cdot)$, then PCA is used to project $\varphi(x)$ into $P(\varphi(x))$, which is a point in a PCA subspace. Preimage learning is used to find the preimage $\hat{x}$ of x in the original input space from $P(\varphi(x))$.

In order to address the pre-image learning problem, some algorithms have been proposed. Mika et al. [134] proposed an iterative method to determine the preimage by minimizing least square distance error. Kwok and Tsang proposed a Distance Constraint Learning (DCL) method to find preimage by using a similar technique in Multi-Dimensional Scaling (MDS) [110]. In a more recent work, Zheng et al. [229] proposed a weakly supervised penalty strategy for preimage learning in KPCA, however their method needs information for both positive and negative classes. As we are only interested in one-class scenarios, the distance constraint method in [110] was selected with respect to the work described in this paper. We briefly review the method here:

For any two patterns x_i and x_j in the input space, the Euclidean distance $d(x_i, x_j)$ can be easily obtained. Similarly, the feature-space distance $\tilde{d}(\varphi(x_i), \varphi(x_j))$ between their φ-mapped images in the feature space can also be obtained. For many commonly used kernels, such as the Gaussian kernels, there is a simple relationship between the feature-space distance and the input-space distance [206]:

$$\tilde{d}_{ij}^2 = \mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\kappa(d_{ij}^2). \tag{5.10}$$

Therefore,

$$\kappa(d_{ij}^2) = \frac{1}{2}(\mathbf{K}_{ii} + \mathbf{K}_{jj} - \tilde{d}_{ij}^2). \tag{5.11}$$

As κ is invertible, d_{ij}^2 can be obtained if $\tilde{d}_{ij}^2$ is known.

Given a training set has n patterns $X = \{x_1, \ldots, x_n\}$. For a pattern x in the input space, the corresponding $\varphi(x)$ is projected to $P(\varphi(x))$, then for each training pattern x_i in X, $P(\varphi(x))$ will be at a certain distance $\tilde{d}(P(\varphi(x)), \varphi(x_i))$ from $\varphi(x_i)$ in the feature space. This feature-space distance can be obtained by:

$$\tilde{d}^2(P(\varphi(x)), \varphi(x)) = \|P(\varphi(x))\|^2 + \|\varphi(x_i)\|^2 - 2P(\varphi(x))'\varphi(x_i). \tag{5.12}$$

The Eqn.(5.12) can be solved by using Eqn.(5.5) and Eqn.(5.8). Therefore, the input-space distances in Eqn.(5.11) between $P(\varphi(x))$ and each x_i can be obtained now. Denote the input-space distance between $P(\varphi(x))$ and x_i as:

$$\boldsymbol{d}^2 = [d_1^2, d_2^2, \ldots, d_n^2]. \tag{5.13}$$

The location of $\hat{x}$ will be obtained by requiring $d^2(\hat{x}, x_i)$ to be as close to the values in Eqn. (5.13) as possible, i.e.,

$$d^2(\hat{x}, x_i) \simeq d_i^2, \quad i = 1, \ldots, n. \tag{5.14}$$

To this end, in DCL, the training set X is constrained to the n nearest neighbors of x, the least square optimization is used to obtain $\hat{x}$.

Construction of One-Class KPCA Ensemble

PCA is a simple method whereby a model for the distribution of training data can be generated. For linear distributions, PCA can be used, however many real world problems are nonlinear. Methods like Gaussian Mixture Models and auto-associative neural networks have been used for nonlinear problems. These methods, however, need to solve a nonlinear optimization problem and are thus prone to local minima and sensitive to the initialization [79]. Kernel PCA was chosen to model the non-linear distribution of training samples. KPCA runs PCA in the high dimensional feature space through the nonlinearity of the kernel, which allows for a refinement in the description of the patterns of interest.

Given an image set of m classes, the proposed one-class KPCA ensemble is built as follows: (i) for each image category, n types image features are extracted; (ii) a KPCA model will be trained for each individual type of extracted features; and therefore (iii) for each image class, n KPCA models will be constructed. For a m-class problem, there will be $m \times n$ KPCA models in the ensemble. The construction of the proposed one-class KPCA ensemble is illustrated in Fig. 5.3.

Multiclass Prediction Using an Ensemble of One-Class KPCA Models

Our classification scheme is designed to produce a reliable prediction for unlabeled images. Classification confidence score is used to describe the probability of the image belonging to each class. The confidence score can provide a quantitative measure of the predictions produced by KPCA models. To disambiguate the competing predictions, a reject option is proposed to evaluate the combined classification result and determines if an unlabeled image should be classified or rejected and passed on to the next stage.

Given an unlabeled image x with n extracted features $F = \{f_1, f_2, \ldots, f_n\}$, let $KPCA_i^j$ represent the KPCA model belonging to class i and trained from the j-th feature f_j, where $i \in \{1 \ldots m\}$ is the class label and $j \in \{1 \ldots n\}$ is the feature label.

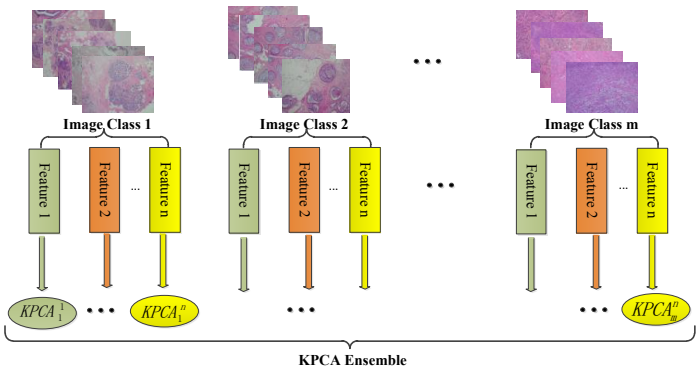


Figure 5.3: Construction of one-class KPCA ensemble from different image feature sets, $KPCA_i^j$ represents the KPCA model trained from the jth image feature of class i.

For classification, each image feature $f_j \in F$ will be reconstructed by all the KPCA models trained from the jth feature. For example, f_1 will be reconstructed by the models $KPCA_i^1, i = 1, \ldots, m$, each of these m KPCA models belongs to one image class. Denote the reconstruction of feature f_j as $f'_j = \{f'^1_j, f'^2_j, \ldots, f'^m_j\}$, we simply use the squared distance D_j between f_j and f'_j as the reconstruction error, thus:

$$D_j = [d_j^1, d_j^2, \ldots, d_j^m], \tag{5.15}$$

where $d_i^j = \|f_j - f'^i_j\|^2, i = 1, \ldots, m$. In the same way, all the features in F will be reconstructed, thus a distance matrix D is obtained, which has the dimensions $n \times m$, where n is the number of KPCA models used for the reconstruction, and m is the number of image classes. Each row of D represents the reconstruction errors of a feature in F by m KPCA models from each class.

$$D = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{bmatrix} = \begin{pmatrix} d_1^1 & d_1^2 & \cdots & d_1^m \\ d_2^1 & d_2^2 & \cdots & d_2^m \\ \vdots & \vdots & \cdots & \vdots \\ d_n^1 & d_n^2 & \cdots & d_n^m \end{pmatrix} \tag{5.16}$$

Note that each column in D represents the reconstruction errors of F using the KPCA models from the same class, these values provide a measure of how x is described by the models from one class. We try to find the KPCA models from one class which give the minimum reconstruction error, this indeed is a 1-nearest neighbor search, as we wish to find the best reconstruction preimage of x in m preimages. Such a distance measure can improve the speed of the classification, moreover, it is also in line with the ideas in metric multidimensional scaling, in which smaller dissimilarities are given more weight, and in locally linear embedding, where only the local neighborhood structure needs to be preserved [110].

In order to combine the reconstruction errors from the KPCA models belonging to the same class, the reconstruction errors in D are normalized using Eqn. (5.17):

$$\tilde{d}_i^j = exp(-d_i^j/s), \tag{5.17}$$

which models a Gaussian distribution from the square distance. The scale parameter s can be fitted to the distribution of d_i^j. Moreover, Eqn. (5.17) has the feature that the scaled value is always bounded between 0 and 1.

The normalized reconstruction errors are then combined to produce the Confidence Scores (CS) of x classified to each class. Let $CS = \{cs_1, cs_2, \ldots, cs_m\}$ denote the confidence scores for x with respect to each image class. The confidence scores are computed from the distance matrix $\tilde{D}$ using a variant of the product rule [99] in Eqn. (5.18):

$$cs_k(x) = \frac{\prod_k P_k(x|w_T)}{\prod_k P_k(x|w_T) + \prod_k P_k(x|w_O)}, \tag{5.18}$$

where k is the number of the combined classifiers. $P_k(x|w_T)$ is the probabilities of classifying x into the target class obtained from k classifiers, and $P_k(x|w_O)$ represents the probabilities of x belonging to the outlier class. In [190], the authors investigated different mechanisms for combining one-class classifiers, their results showed that the "product rule" outperforms other combining mechanisms for one-class classifiers.

As noted in [190, 99], when using the product combining rule, $P_k(x|w_T)$ should be available and a distance should be transformed to a "resemblance" by some heuristic mapping as in Eqn. (5.17). However, when an image feature is reconstructed by a number of KPCA models from different image classes, some models will give big reconstruction errors, which will become relatively small, approaching 0 after the mapping of Eqn. (5.17), this makes the item $\prod_k P_k(x|w_O)$ in Eqn. (5.18) meaningless. Therefore, we propose to use a variant of the product combining rule in (5.18). Instead of using the mapping values from all KPCA models, for the KPCA models trained by the same type of image feature, only the model that gives the biggest mapping value will be chosen to produce $\prod_k P_k(x|w_O)$. The proposed product combining rule can be described as:

$$cs_k(x) = \frac{\prod_k P_k(x|w_T)}{\prod_k P_k(x|w_T) + \prod_k \max P_k(x|w_O)}. \tag{5.19}$$

This maximum value selection procedure is illustrated in Fig. 5.4 by a simple example. In Fig. 5.4, there is a 4-class classification task (I, II, III, IV in the figure), four types of features are extracted from image x. For one type of image feature, there are four trained KPCA models, each from a different class, giving four reconstruction results for the same feature of x (one row in matrix $\tilde{D}$). If we consider class I as the 'target' class (first column in the figure), the four values in the first column are used to

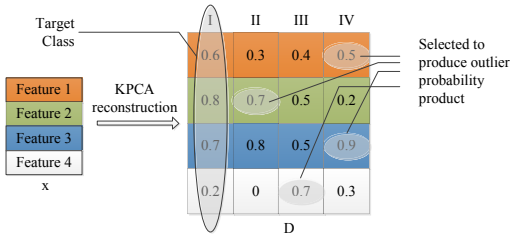


Figure 5.4: Illustration of KPCA model selection to produce outlier probability product.

produce the item $\prod_k P_k(x|w_T)$ in (5.19). The other three column of values are deemed as the outlier probabilities produced by the KPCA models from the other three classes. The proposed combining rule selects the maximum mapping value from each row to produce the outlier probability product $\prod_k P_k(x|w_O)$.

The proposed combining rule is in line with the basic idea of one-class classification, as in the one-class scenario one only needs to know if a pattern should be assigned to the target class or to the outlier class. If one or more outlier models is able to produce a high outlier probability product, the current target class should be doubted. Moreover, by combining the outliers value from different feature-derived models, the diversity of the ensemble will be improved, which is an important factor to make an ensemble learning method successful [108].

To classify an unlabeled image, each class will be regarded as the target class in turn, using the proposed product combining rule, a classification confidence score can be obtained for assigning x to each class. The procedure of obtaining the confidence scores is described in Algorithm 3.

Once the CS set has been obtained, the decision to classify or reject can be made. We first give two parameters that will be used later:

Definition 1 (Top Confidence Score)

$$CF_T = \max\{CS\}$$

Definition 2 (Class Confidence Margin)

$$CF_M = CF_T - \max\{CS - \{CF_T\}\} \tag{5.20}$$

Although CF_T is the highest confidence score from the combination of m KPCA models, it is suggested that using only CF_T for classification is not sufficiently accurate. In [65], the authors demonstrated that during classification, the correct predictions tend to have both high CF_T and CF_M, whereas the wrong predictions may have high CF_T but smaller CF_M. Therefore, to use both CF_T and CF_M as classification measures can decrease the appearance of wrong predictions.

Algorithm 3 Calculation of confidence scores for classifying x into each class

Input:

$M = \{1, \ldots, m\}$: Class label set
D: Distance matrix
$i = 1 \ldots m$: Class label index
$j = 1 \ldots n$: Image feature index
I: Target class label set
L: Outlier class label set
PT_i: Product of target class probabilities
PO_i: Product of outlier class probabilites

Output:

$CS = \{cs_1, cs_2, \ldots, cs_m\}$: Confidence scores for assigning x into each class

1: $CS = \emptyset$;
2: **for** $(i = 1;\ i \leq m;\ i++)$ **do**
3: $PT_i = 1; PO_i = 1; I \leftarrow i; L = M - I$;
4: **for** $(j = 1;\ j \leq n;\ j++)$ **do**
5: $PT_i = PT_i \times \hat{d}_j^i$;
6: $PO_i = PO_i \times \max\{\hat{d}_j^L\}$;
7: **end for**
8: $cs_i = (PT_i / PT_i + PO_i)$;
9: $CS = CS \cup cs_i$;
10: **end for**
11: **return** CS

5.3.2 Reject Option for Classification

As already noted, in order to obtain a reliable classification system, the rejection option is used here. The optimal classification rule with reject option was defined by Chow [30]. 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 maximum posterior probabilities can be identified as a measure of classification reliability. Such a rule can be expressed as:

$$f(x) = \begin{cases} argmax_{C_i}(p(C_i|x)) & \text{if } max_{C_i}\ (p(C_i|x)) \geq t \\ \\ reject & \text{if } \forall_i\ p(C_i|x) < t \end{cases} \tag{5.21}$$

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

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

$$p(reject) = \int_{reject} p(x)dx = p(max(p(C_i|x)) < t). \tag{5.22}$$

In Chow's theory, an optimal classifier can be found only if the true posterior proba-

bilities 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, when the two confidence parameters CF and CF_M are obtained, two thresholds t_1 and t_2 are selected to control the reliability-rejection trade-off. An image will be rejected if its confidence $CF_T(x)$ and $CF_M(x)$ cannot satisfy the rejection rule in Eqn. (5.23):

$$CF_T(x) \geq t_1 \; and \; CF_M(x) \geq t_2. \tag{5.23}$$

We will not pursue the optimal error-reject trade-off, as different image sets will have different optimal rejection thresholds. The value of t_1 will be set as a fixed number and the value of t_2 for each class is determined by a simple rule so that the selected threshold t_2 results in the max difference between classification reliability and rejection rate.

5.3.3 Random Subspace Ensemble of One-versus-All SVMs

The rejected samples from the KPCA ensemble in Stage 1 will be handled by the second ensemble, which is a Random Subspace ensemble of one-versus-all SVMs. The Random Subspace (RS) method [78] is often quoted as an efficient way of combining the results of a number of classifiers. 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 our scheme, the multiclass classification problems with K classes are decomposed into K independent two-class problems (the *one-versus-all* approach where each classifier classifies records as belonging or not belonging to a class). The multiclass classification task can then be conducted based on the outputs of the binary SVMs. Denote the output of a SVM as $\rho(x)$ for an unlabeled pattern x, to estimate class posteriors from the SVM's output, a mapping can be implemented using:

$$P(y = +1|\mathbf{x}) = \frac{1}{1 + \exp(\mathrm{a}\rho(\mathbf{x}) + \mathrm{b})} \tag{5.24}$$

where the class labels are denoted as $y = +1, -1$, while a and b are constant terms to be defined on the basis of the 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 $\rho(x)$.

Therefore, K binary SVM classifiers are constructed for K different image classes. We refer to such a K collection of binary SVMs as an *expert* to avoid the confusion with *ensemble*. The ith SVM output function P_i is trained taking the examples from the i-th

class as positive and the examples from all other classes as negative. In other words, 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. For a new sample x, the corresponding SVM assigns it to the class with the largest value of P_i as follows:

$$Class = \arg\max \ P_i, \quad i = 1, \ldots, n \tag{5.25}$$

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 2 ensemble, trained with randomly chosen subsets of all available features (*i.e.* random subspaces) following the Random Subspace strategy. In the random subspace strategy, 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 submitting 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 SVM 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_3 of the experts are agreed on the label, otherwise, the sample is rejected, the threshold t_3 can be decided in advance. For example, a simple rule as follows can be used to decide the value of t_3:

$$t_3 \geq \begin{cases} \frac{M}{2} + 1 & \text{if } M \textit{ is even} \\ \frac{M+1}{2} & \text{if } M \textit{ is odd.} \end{cases} \tag{5.26}$$

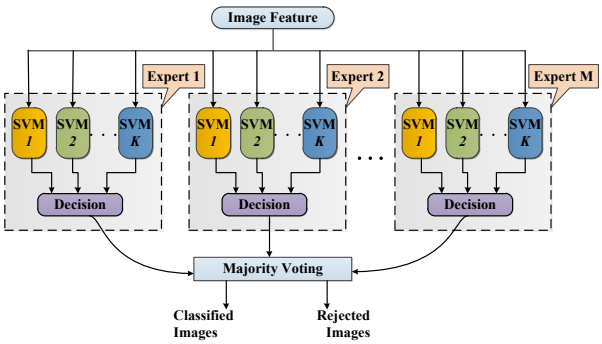


Figure 5.5: SVM ensemble with rejection option in Stage 2, which consists of a set of binary SVMs (experts).

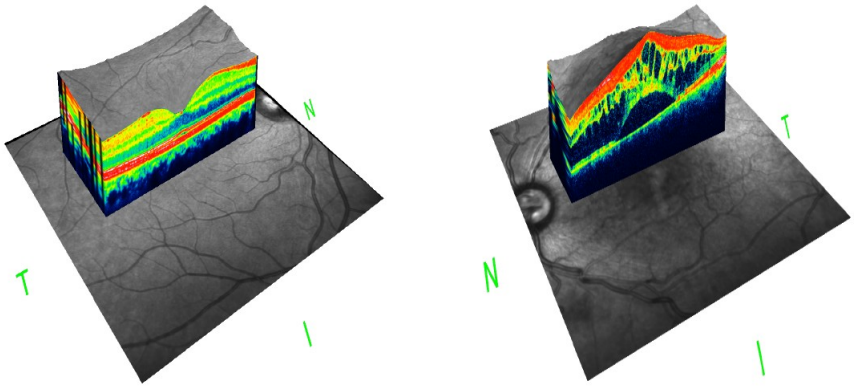
Since there can be more than two classes, the combined decision is deemed to be correct when a majority of the experts are correct, but wrong when a majority of the decisions are wrong. Obviously, t_3 is a tunable threshold that controls the rejection

rate, and we use t_3 to relate the consensus degree from the majority voting to the confidence measure, and abstain from classifying ambiguous samples. Fig. 5.5 further explains the principle of the SVM ensemble in stage 2.

5.4 Experiments and Results

The effectiveness of the proposed method is illustrated using a biopsy breast cancer benchmark image set and a 3D OCT retinal image set, the details of the biopsy image set are introduced in Section 3.3.2. The feature extraction methods of the biopsy images are introduced in Section 3.4.

The 3D OCT retinal image set was collected at the Royal Hospital of University of Liverpool [5], the image set contains 140 volumetric OCT images, in which 68 images from normal eyes and the remainders are from eyes have Age-related Macular Degeneration (AMD). Fig 5.6 shows the example images.



(a) A 3D OCT for a normal eye (b) A 3D OCT for an AMD eye

Figure 5.6: Examples of two 3D OCT images showing the difference between a "normal" and an AMD retina [4].

The OCT images are preprocessed by using the Split Bregman Isotropic Total Variation algorithm with a least-squares approach. The preprocessing step has two targets: (i) identification and extraction of a Volume Of Interest (VOI) which also results in noise removal, and (ii) flattening of the retina as appropriate. The example images after preprocessing can be seen in Fig. 5.7.

Section 5.4.1 introduces our experimental setup and the evaluation methods used in our experiments. The six commonly used one-class classifiers are compared with two synthetic datasets in section 5.4.2. Using the extracted image features of the biopsy image set, the effectiveness of combining Kernel PCAs is illustrated in section 5.4.3. Then in Section 5.4.4, the performance of the proposed system is also evaluated and compared on the 3D OCT retinal image set.

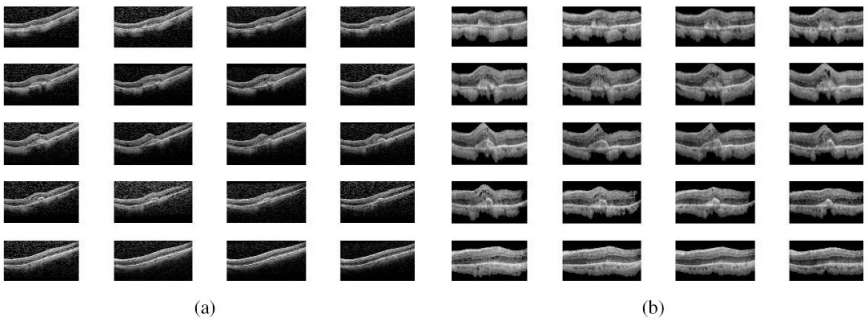


Figure 5.7: Examples of OCT images. (a) Before preprocessing. (b) After preprocessing. [4]

5.4.1 Experimental Setup and Performance Evaluation Methods

MATLAB 7.0 was used to implement the proposed process together with the Gaussian kernel $k(x, y) = exp(-\|x - y\|^2/2\sigma^2)$. Other types of kernels could have been used, however.

Unless other wise stated 10-fold cross validation was used, all the results are averages of 10 runs of the 10-fold cross validation. The following measures are used to evaluate the proposed cascade method:

- Recognition rate (RR) = number of correctly recognized images / (number of testing images - number of rejected images).
- Rejection rate (RejR) = number of rejected images /number of testing images.
- Reliability (RE) = (number of correctly recognized images + number of rejected images)/ number of testing images.
- Error rate (ER): = 100% - reliability.
- ROC: Receiver Operating Characteristic graph.
- AUC: Area under an ROC curve.

5.4.2 Comparison among Different One-Class Classifiers

In this section, we use two synthetic datasets to evaluate the kernel PCA classifier by comparing the decision boundaries of KPCA with five other commonly used one-class classifiers: (i) PCA, (ii) MoG (Mixture of Gaussians with 2 components), (iii) k-means, (iv) SVDD, (v) Parzen.

Fig. 5.8 shows the classification boundaries of the compared classifiers on a banana-shaped dataset which has 120 data points. For kernel PCA, MoG, SVDD and Parzen density estimation, the width of the Gaussian kernel is set to $\sigma = 4$. The number of

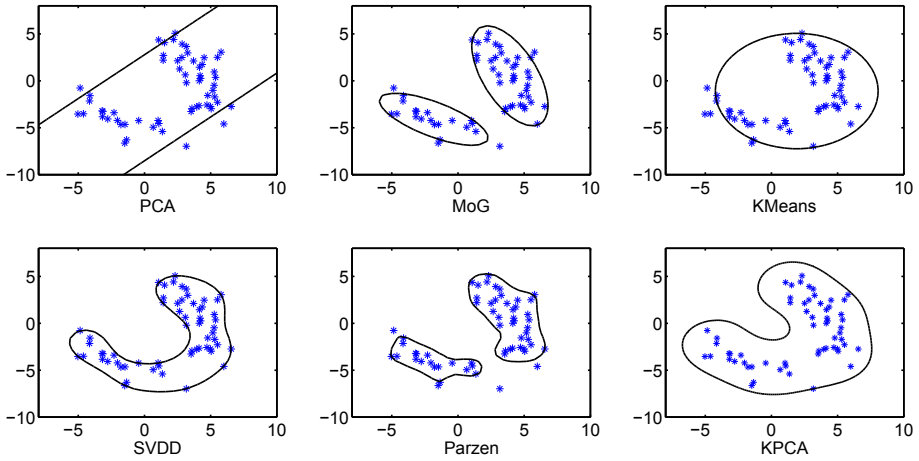


Figure 5.8: Classification Boundaries of Different One-Class Classifiers on Banana dataset.

eigenvectors used for reconstruction in PCA and KPCA is set to $n = 40$. A decision threshold of 0.1 was selected for all the classifiers, which identifies 10% of the data as outliers during training in order to improve the generalities of the classifiers. As can be seen in the figure, the PCA, MoG, k-means and Parzen density estimation are unable to describe the distribution. The KPCA and SVDD provide a better description of the data, however, SVDD does not generalize well since the decision boundary contains irregularities of the data distribution.

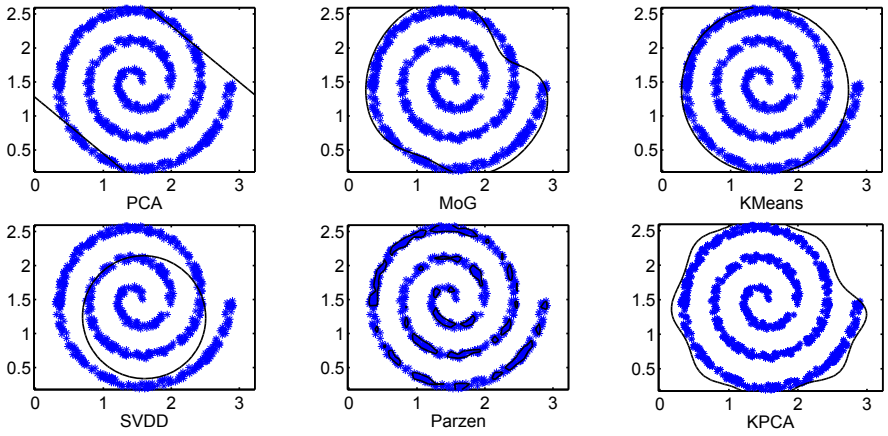


Figure 5.9: Classification Boundaries of Different One-Class Classifiers on Spiral dataset.

To test how kernel PCAs can cope with more complex data distributions, a spiral distribution which contains 700 data points [79] is used. In Fig. 5.9, all the classifiers

use the same parameters as in the banana data test. Although in Fig. 5.9 all the classifiers fail to describe the distribution, however, when the width of the Gaussian kernel is changed to a smaller value, KPCA can describe the distribution well. The SVDD still cannot improve the decision boundary with a smaller σ, we compare the operation of KPCA and SVDD in Fig. 5.10 with $\sigma = 0.25$.

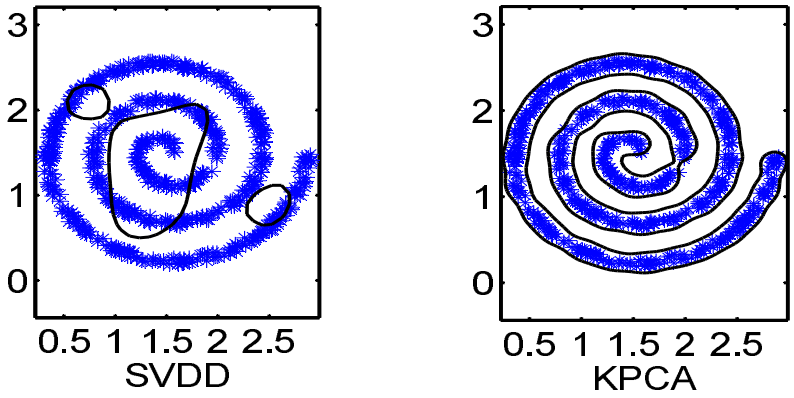


Figure 5.10: Classification Boundary of KPCA and SVDD on Spiral dataset with $\sigma = 0.25$.

5.4.3 Results on Breast Cancer Biopsy Image Set

The KPCA ensemble evaluation using the biopsy image data is reported in this section. Three types of image features were extracted, therefore for each image class three Kernel PCAs were built with respect to each type of image feature. The recognition rates of using these KPCAs individually are listed in column 2 to column 4 of Table 5.1, where CvletK, GLCMK and LBPK represent KPCA models trained from Curvlets, GLCM and LBP, respectively. The results of combining all KPCA models are listed in the last two columns of Table 5.1. Column 5 gives the results from the original combining product rule introduced in Eqn. (5.18). The results from the proposed product combining rule (Eqn. (5.19)) are listed in the sixth column. The parameters of KPCAs were set to $\sigma = 4$ and $n = 40$.

Table 5.1: Recognition rate (%) for the biopsy image data from individual KPCAs and the combined model.

Image Class	CvletK	GLCMK	LBPK	Original combining rule	Proposed combining rule
Normal	70.10	67.70	71.40	69.25	92.70
Insitu	76.50	72.58	81.83	74.47	93.78
Invasive	77.71	68.65	85.57	75.22	90.35

Note that the results in Table 5.1 were obtained without rejection, each image is directly assigned to the class with the Top Confidence Score (CF_T). From Table 5.1 one

can see that by using the proposed product combining rule, the classification accuracies of all the image classes have been improved. This illustrates that by combining one-class classifiers trained from different features can improve the classification performance, which is in accordance with the observation in [190]. For comparison, the other one-class classifiers are also used as the base classifier of the ensemble in Stage 1, using the same combining rule, the classification results are listed in Table 5.2.

Table 5.2: Recognition rate (%) for the biopsy image data from different one-class classifier ensembles. The kernel widths for KPCA and SVDD were set to $\sigma = 4$. The number of principal components for KPCA and PCA were set to $n = 40$.

Image Class	PCA	MoG	KMeans	SVDD	Parzen	KPCA
Normal	85.17	82.12	80.12	85.56	84.54	92.70
Insitu	87.33	84.67	83.46	87.22	81.26	93.78
Invasive	82.56	81.88	79.65	84.67	83.23	90.35

In the next expriment, the rejection option in Eqn. (5.23) combined with the two confidence scores defined in Eqn. (5.14) are used, the experimental results showed that with the rejection option and the control of the reliability-rejection tradeoff, the proposed cascade system obtained promising results using the biopsy image data.

Without losing generality and simplicity, in the following experiments, the top confidence score (CF_T) of each image class was empirically set to 0.5. In order to investigate the effectiveness of the rejection rule, the class confidence margins (CF_M) were increased from 0 to 0.4 in steps of 0.02. Using different values for the CF_M threshold, the performance of the first stage (KPCA ensemble) in the cascade system is shown in Fig. 5.11.

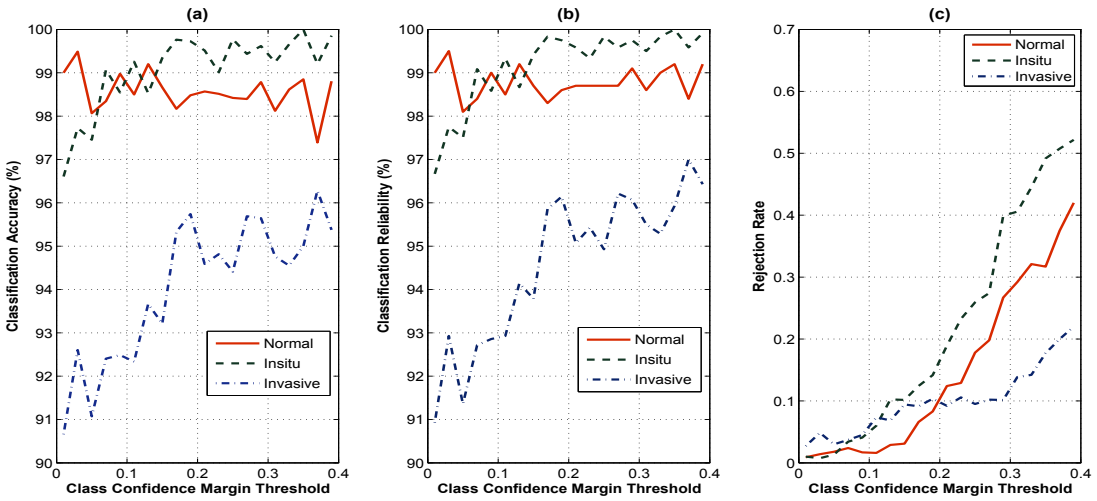


Figure 5.11: Classification performance of KPCA ensemble in Stage 1 with different CF_M threshold values.

From Fig. 5.11 (a) it can be observed that, as the rejection threshold CF_M was increased, the classification accuracies (recognition rates) of the three image classes improved. However, for the class of "normal", the recognition rate reaches a peak of 99.48% at a rejection threshold of 0.02, this indicates that the ambiguous images in this class produce small class confidence margins, a small threshold can reject these ambiguous images and improve the recognition rate of the class. The highest classification reliability for the "normal" class was 99.52% obtained with $CF_M = 0.02$. The rejection rate for the "normal" class in this case was only 1.4% for the top recognition rate and reliability.

The recognition rates of other two image classes ("insitu" and "invasive") reached the highest points at 100% and 96.28%, respectively, using a rejection threshold of 0.35 for both classes. The classification reliabilities of these two classes were also the best, obtained 100% for the "insitu" class and 97.00% for the "invasive" class (Fig. 5.11 (b)). It can be seen from Fig. 5.11 (c) that when the rejection threshold is 0.35, for the "insitu" class, the rejection rate is 49.17% and the rejection rate for the "invasive" class is 17.71%. This means that to reach a better performance in Stage 1 for these two classes, more images need to be rejected with respect to the second stage.

However, the side effect of a high rejection rate is that it has the potential to enhance the error rate of the next stage. Therefore, a simple rule was used here to determine the rejection threshold t for each class, the selected t was the threshold that gave the maximum difference between the classification reliability and rejection rate, namely, we chose the rejection threshold t that gave the maximum of $|RE - RejR|$, which can be written as:

$$t = argmax|RE_t - RejR_t|. \tag{5.27}$$

This simple rule guarantees that the selected threshold produced a high classification reliability with a small rejection rate. With this rule, the best thresholds for the three image classes in Stage 1 are listed in Table 5.3 (Column TH), the corresponding classification performances are also listed (recognition rate, reliability, rejection rate, error rate).

Table 5.3: Best classification performance for the biopsy image data for the KPCA ensemble, where RR, RE, RejR and ER represent recognition rate, reliability, rejection rate and error rate. TH represents the rejection threshold that produced the results.

Image Class	RR (%)	RE (%)	RejR (%)	ER (%)	TH
Normal	99.48	99.52	1.40	0.48	0.02
Insitu	99.76	99.83	12.42	0.17	0.17
Invasive	96.28	97.00	17.71	3.00	0.35

The images rejected by the KPCA ensemble will be further classified or rejected in Stage 2, which is a "one-versus-all" SVM ensemble. The library for support vector

machines, LIBSVM[1] was used for the experiments. The parameter σ that defines the spread of the radial function set to 5.0 and the parameter C that defines the trade-off between the classifier accuracy and the margin 3.0.

For the SVM ensemble training, the three kinds of image feature vectors were combined together forming a single feature vector for each image. Two important parameters for the Random Subspace ensemble are the ensemble size L and the cardinality of the feature vectors M (the size of the randomly chosen subsets of all available features, thus the random subspace). A "rule of thumb" has been put forward with respect to the fMRI data classification problem [107], 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. However, in our previous work [223], it was observed that this rule does not work well with the biopsy image classification. Based on our previous research, $M = \frac{4n}{5}$ is used for the random subspace training of SVM ensemble. In [223], it was found that a bigger ensemble size ($L > \frac{n}{10}$) may bring better classification performance, however the big ensemble sizes also bring a heavy computational cost. With respect to the work in this paper, we used the rule $L = \frac{n}{10}$, the ensemble size was set as $L = 65$ for evaluating our system, as the dimension of the combined feature was $n = 666$.

Majority voting was used in Stage 2 to control the reliability-rejection tradeoff. In combining the decisions from the M SVM experts (Figure 5.5), a sample is assigned the class for which at least t of the experts are agreed on the label, otherwise, the sample is rejected. For evaluating the error-tradeoff of the second stage, the threshold t was increased from $t = 32$ to $t = 65$ in steps of 3. The classification performance of Stage 2 using different thresholds is shown in Fig. 5.12. As the second stage is a multiclass classifier, the classification results were obtained using all the rejected images from Stage 1.

From Fig. 5.12 (a) it can be seen that as the rejection threshold increases from 32 to 50, the recognition rate and reliability values improve. At a threshold of 50, the recognition rate and reliability of the second stage reach their peak values of 97.88% and 98.20% respectively. As the rejection threshold continues to increase, it can be seen that there is no further improvement in classification performance. At the same time, the high thresholds bring high rejection rates, as shown in Fig. 5.12 (b), when the threshold is 65, the rejection rate of the second stage is 37.61%. Therefore, the threshold value 50 was selected as the optimal threshold value for the SVM ensemble, as it produced the maximum difference between reliability and rejection rate. Table 5.4 lists the classification results for the three image classes using a rejection threshold of 50.

In Table 5.4, one can see that when using a threshold of 50, for the classes "normal"

[1] www.csie.ntu.edu.tw/ cjlin/libsvm

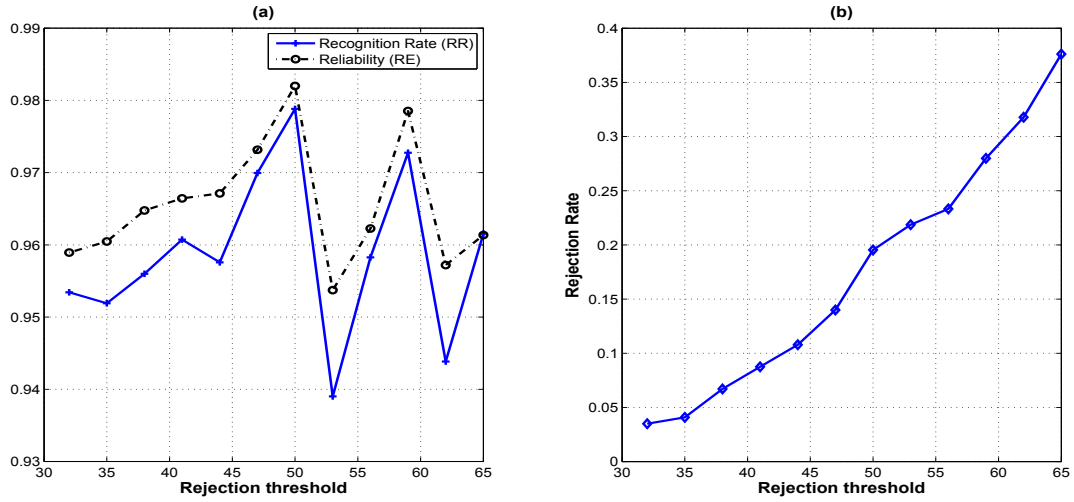


Figure 5.12: Classification performance of SVM ensemble in stage 2 with different rejection threshold values.

Table 5.4: Classification performance of Stage 2 on the biopsy image set

Image Class	RR (%)	RE (%)	RejR (%)	ER (%)	TH
Normal	1	1	0	0	50
Insitu	93.65	94.60	12.08	5.40	50
Invasive	1	1	27.50	0	50

and "invasive", the rejected images from Stage 1 can be correctly classified with small rejected numbers. Under the selected optimal thresholds for Stage 1 and Stage 2, the overall classification performance of the proposed cascade system is listed in Table 5.5. The classification confusion matrix is presented in Table 5.6.

Table 5.5: Overall classification performance for the biopsy image data of the proposed cascade system

Image Class	RR (%)	RE (%)	RejR (%)	ER (%)
Normal	99.50	99.50	0	0.50
Insitu	98.33	99.83	1.5	0.17
Invasive	97.57	99.43	3.5	0.57
Overall	98.36	99.58	1.86	0.42

The results with respect to the evaluation image dataset obtained using other methods were also considered. In [16], 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. An error rate of 1.25% and rejection rate of 1.94% were reported in [223], which also used a cascade classification scheme, however our proposed method produces an error rate of 0.42% with a smaller rejection rate of 1.86%.

With respect to the comparison of a variety of one-class classifiers, the classifiers from Section 5.4.2 were used as the base classifiers for the ensemble of Stage 1. The Receiver Operating Characteristics (ROC) curves obtained using different one-class classifiers are shown in Fig. 5.13. The Areas Under the ROC curves (AUC), for the compared classifiers, are listed in Table 5.7, the KPCA ensemble gives the best result.

5.4.4 Results on the 3D OCT Retinal Image Set

The 3D OCT retinal image contains 140 images, in which 68 are normal eyes and the remainders are AMD (Age-related Macular Degeneration). To further evaluate the proposed method on imbalanced data problem, in each run of the experiments, only 40 images in the normal class were randomly chosen for classifier training. As the images are three-dimensional, following the work in [4], three types image features were used for image description: Local Binary Patterns of Three Orthogonal Planes (LBP-TOP), Local Phase Quantization (LPQ) and Multi-Scale Spatial Pyramid (MSSP).

Table 5.8 presents the classification results of KPCA models trained by individual

Table 5.6: Averaged confusion matrix with overall rejection rate 1.86% (%)

	insitu	normal	invasive	rejected
normal	99.50	0.14	0.36	0
insitu	0.37	98.33	1.30	1.5
invasive	1.26	1.17	97.57	3.5

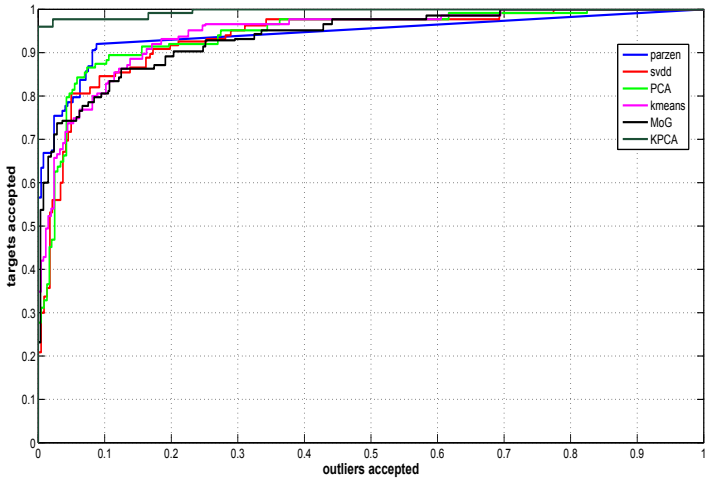


Figure 5.13: Receiver operating characteristics curves of different one-class classifiers used as the base classifiers for the ensemble of stage 1.

Table 5.7: AUC of different one-class classifiers used as the base classifier for the ensemble of stage 1.

	Parzen	SVDD	PCA	Kmeans	MoG	KPCA
AUC	94.30	93.61	94.19	94.28	93.67	99.53

features and the combined feature. From Table 5.8 one can see that by using the proposed product combining rule in Eqn. (5.19), the classification accuracies of all the image classes have been improved.

Table 5.8: Recognition rate (%) for the 3D OCT retinal image data from individual KPCAs and the combined model.

Image Class	LPQ	LBP-TOP	MSSP	Original combining rule	Proposed combining rule
Normal	86.20	88.45	85.56	78.83	91.30
AMD	86.50	87.69	85.83	74.67	90.22

Apply the reject option on the KPCA ensemble in stage 1, using the rejection threshold selection rule in Eqn. 5.27, the classification performance of stage 1 on the 3D OCT retinal images can be seen in Table 5.9.

As for the second stage, we simply used the same rejection threshold for the biopsy image set, under the rejection threshold 50, the performance of stage 2 on 3D OCT retinal image set is listed in Table 5.10. The overall performance on the image set is presented in Table 5.11.

The Receiver Operating Characteristics (ROC) curves obtained using different one-class classifiers are shown in Fig. 5.14. For comparison, the results from a recent publication [4] using the same 3D OCT dataset and the results of the proposed method

Table 5.9: Best classification performance for the 3D OCT retinal image data for the KPCA ensemble, where RR, RE, RejR and ER represent recognition rate, reliability, rejection rate and error rate. TH represents the rejection threshold that produced the results.

Image Class	RR (%)	RE (%)	RejR (%)	ER (%)	TH
Normal	94.35	96.45	7.54	3.55	0.07
AMD	93.56	95.77	10.86	4.23	0.12

Table 5.10: Classification performance of stage 2 on the 3D OCT retinal image set.

Image Class	RR (%)	RE (%)	RejR (%)	ER (%)	TH
Normal	93.24	94.36	13.68	5.64	50
AMD	92.11	93.15	16.66	6.85	50

are listed in Table 5.12.

5.5 Conclusion

In this chapter, a reliable classification scheme based on the serial fusion of a one-class KPCA model ensemble together with a random subspace SVM ensemble has been proposed for medical image classification. 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 one-class KPCA models trained using different image features from each image class, while the second ensemble comprises SVMs. During classifier generation, randomly sampled subsets of features, following the Random Subspace procedure, were used. For both of the ensembles the reject option was implemented using a confidence threshold.

The effectiveness of the proposed cascade classification scheme was verified using a breast cancer biopsy image dataset and a 3D OCT retinal image set. The two-stage ensemble cascade classification scheme obtained high classification accuracies and simultaneously guaranteed high classification reliabilities with small rejection rates. The proposed cascade system obtained a 98.36% classification accuracy and a 99.58% classification reliability on the biopsy image set. Compared with the state-of-the-art result

Table 5.11: Overall classification performance on the 3D OCT retinal image set.

Image Class	RR (%)	RE (%)	RejR (%)	ER (%)
Normal	94.78	95.15	0.38	4.85
AMD	94.33	94.67	0.33	5.23
Overall	94.56	94.91	0.36	5.04

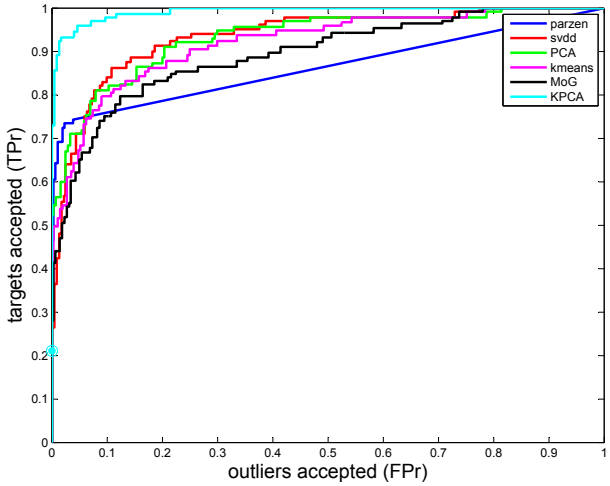


Figure 5.14: Receiver operating characteristics curves for 3D OCT retinal image set with different one-class classifiers used as the base classifiers for the ensemble of stage 1.

Table 5.12: AUC and classification accuracy comparison of 3D OCT retinal image set

	results in [4]	proposed method
AUC	94.40	95.33
Classification accuracy	91.50	94.56

on the same image set, the proposed method obtained a 4.66% improvement on the classification accuracy. For the 3D OCT retina image set, a classification accuracy of 94.40% was obtained using the proposed cascade method, which achieves a 2.9% improvement compared to the published result. 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.

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 effort therefore needs to be put into detecting ROIs from biopsy images. In this work, the parameters for the cascade system, such as ensemble size and rejection threshold, were decided empirically; this may not produce 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.

Chapter 6

Conclusions and Future Work

In this thesis, the problem of biomedical image classification is investigated. The random subspace method for classifier ensemble is used for combining different classifiers trained by multiple image features. A new cascade classification scheme based on reject option is developed to improve the classification accuracy and reliability for medical image classification problems. In order to address the problem of imbalanced data problem in many medical image diagnosis applications, a new ensemble of one-class classifiers is developed, where the reject option is also included to construct a cascade classifier. The classification schemes proposed in this thesis can be summarized as follows:

- A random subspace ensemble of neural networks is proposed to classify microscope images. Using a combination of three image descriptors, namely curvelet transform, gray level co-occurrence matrix and completed local binary patterns, the designed paradigm is well-suited to the characteristics of microscopic image data. Experiments on the benchmark RNAi datasets showed that the random subspace MLP ensemble method achieved higher classification accuracies ($\sim 87.1\%$). Compared to the published result 82%, a 4.9% improvement on the classification accuracy was obtained. The classification results of other three groups of microscopy image data sets using random subspace MLP also support the effectiveness of the proposed method. The random subspace MLP ensemble obtained 86.6% classification accuracy on the 2D Hela dataset, and 93.7% on the CHO dataset, providing the improvements of 0.7% and 2.6% on the classification accuracy, respectively. A classification accuracy of 95.22% was obtained from the proposed ensemble method on the biopsy image sets, which obtains an 1.82% improvement on the published result on the same image sets [113].

- A reliable classification scheme based on cascaded 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 mis-

classifications so as to ensure high classification reliability. The proposed cascade 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 reject 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 two-stage ensemble cascade classification scheme resulted in a high classification accuracy (99.25%) and simultaneously guaranteed a high classification reliability (97.65%) with a small rejection rate (1.94%). We have observed a 5.6% improvement on the classification accuracy compared with the best published result [16]. Moreover, the cascade architecture provides a mechanism to balance between classification accuracy and rejection rate.

- A novel classification scheme based on the serial fusion of a one-class KPCA model ensemble together with a random subspace SVM ensemble has been proposed for medical image classification. The first stage ensemble consists of one-class KPCA models trained using different image features from each image class, while the second ensemble comprises SVMs. During ensemble construction, randomly sampled subsets of features were used following the Random Subspace procedure. For both of the ensembles the reject option was implemented using a confidence threshold. The effectiveness of the proposed cascade classification scheme was verified using a breast cancer biopsy image dataset and a 3D OCT retinal image set. The proposed cascade system obtained a 98.36% classification accuracy and a 99.58% classification reliability on the biopsy image set. Compared with the state-of-the-art result on the same image set [16], the proposed method obtained a 4.66% improvement on the classification accuracy. For the 3D OCT retina image set, a classification accuracy of 94.40% was obtained using the proposed cascade method, which achieves a 2.9% improvement compared to the published result [4].

To sum up, research effort has been taken on developing and implementing new algorithms to solve the biomedical image classification problem, particularly for microscope images. It has been verified from our experiments that using classifier ensemble can improve the classification performance. The random subspace based ensemble led to superior results over popular ensemble strategies. The proposed two-stage classification schemes composed by different classifier ensembles further enhance the

classification accuracy. The use of reject options in the cascade systems can simultaneously guarantee high accuracy and reliability of the classification. By investigating the error-reject trade-offs, appropriate rejection thresholds were selected for different classification tasks, this resulted in high classification accuracy and reliability under small rejection rates.

Although the proposed methods achieved promising results with respect to the classification of biomedical images, there are several aspects that can 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 (RoI) that contain the most important tissues in biopsy scans. Therefor, more effort therefore needs to be put into detecting ROIs from biopsy images.

- In this thesis, the parameters for the cascade system (e.g. ensemble size, rejection threshold) were decided empirically; this may not produce 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.

- The random subspace utilizes different feature subspaces to guarantee the diversity of base classifiers in an ensemble. However, in the current work, the diversity of the proposed systems were not theoretically investigated. In future research, quantitative analysis of ensemble diversity and its effects on the classification performance will be carried out.

- In this thesis, the classification reliability is a measurement for the whole classification systems obtained from all testing samples. However, the classification reliability for a single sample is also important in medical applications, where the accuracy of prediction for any individual patient is more important than the global error of the classification model. In order to guarantee high reliability for each individual sample, some dynamic ensemble generation methods can be incorporated into current schemes to deal with 'ambiguous' samples. Another alternative way is to use transductive inference classifiers [143] as the base classifiers in the ensemble.

Bibliography

[1] Chebira A., Barbotin Y., Jackson C., Merryman T., Srinivasa G., Murphy RF., and Kovacević J. A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, 8(210):1–10, 2007.

[2] Tsymbal A. and Puuronen S. Ensemble feature selection with the simple bayesian classification in medical diagnosis. In *Proceedings of 15th IEEE symposium on Computer-Based Medical Systems*, pages 225–230, 2002.

[3] Gaurav Agarwal, P. V. Pradeep, Vivek Aggarwal, Cheng-Har Yip, and Polly S. Y. Cheung. Spectrum of breast cancer in asian women. *World Journal of Surgery*, 31(5):1031–1041, 2007.

[4] Abdulrahman Albarrak, Frans Coenen, and Yalin Zheng. Age-related macular degeneration identification in volumetric optical coherence tomography using decomposition and local feature extraction. In *Proceedings of 2013 International Conference on Medical Image, Understanding and Analysis*, pages 59–64, 2013.

[5] Abdulrahman Albarrak, Frans Coenen, Yalin Zheng, and Yu W. Volumetric image mining based on decomposition and graph analysis: An application to retinal optical coherence tomography. In *Proc. CINTI 2012, Budapest, Hungary*, pages 263–268, 2012.

[6] R. Arisio, C. Cuccorese, G. Accinelli, M.P. Mano, R. Bordon, and L. Fessia. Role of fine-needle aspiration biopsy in breast lesions: Analysis of a series of 4,110 cases. *Diagnostic Cytopathology*, 18(2):462–467, 1998.

[7] S. Arivazhagan and L. Ganesan. Texture classification using wavelet transform. *Pattern Recognition Letters*, 24:1513–1521, 2003.

[8] Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

[9] Ajay Nagesh Basavanhally, Shridar Ganesan, Shannon Agner, James Peter Monaco can Michael D. Feldman, John E. Tomaszewski, Gyan Bhanot, and

Anant Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Transactions on Biomedical Engineering*, 57(3):642–653, 2010.

[10] C. Bergamini, L.S. Oliveira, A.L. Koerich, and R. Sabourin. Combining different biometric traits with one-class classification. *Signal Processing*, 89:2117–2127, 2009.

[11] A. Bertoni, R. Folgieri, and G. Valentini. *Biological and Artificial Intelligence Environments*. Springer-Verlag, Berlin, 2008.

[12] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is 'nearest neighbor' meaningful? *Lecture Notes in Computer Science*, 540:217–235, 1999.

[13] Michael V. Boland. *Quantitative Description and Automated Classification of Cellular Protein Localization Patterns in Fluorescence Microscope Images of Mammalian Cells*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, July 1999.

[14] L. E. Boucheron. *Object- and Spatial-Level Quantitative Analysis of Multispectral Histopathology Images for Detection and Characterization of Cancer*. PhD thesis, University of California Santa Barbara, Santa Barbara, CA, 2008.

[15] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[16] A. Brook, R. El-Yaniv, E. Isler, R. Kimmel, R. Meir, and D. Peleg. Breast cancer diagnosis from biopsy images using generic features and svms. Technical Report CS-2008-07, Technion-Israel Institute of Technology, Technion City, Haifa 32000, Isreal, 2006.

[17] Gavin Brown. *Diversity in Neural Network Ensembles*. PhD thesis, University of Birmingham, Birmingham, UK, 2004.

[18] Lorenzo Bruzzone and Diego Fernández Prieto. A partially unsupervised cascade classifier for the analysis of multitemporal remote-sensing images. *Pattern Recognition Letters*, 23:1063–1071, 1998.

[19] Leistner C., Saffari A., Roth P. M., and Bischof H. On robustness of on-line boosting - a competitive study. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1362–1369, 2009.

[20] E. Candes, L. Demanet, D. Donoho, and L. Ying. Fast discrete curvelet transforms. *Multiscale Modeling and Simulation*, 5:861–899, 2006.

[21] E. J. Candès and D. L. Donoho. Continuous curvelet transform: I. resolution of the wavefront set. *Appl. Comput. Harmon. Anal.*, 19:162–197, 2003.

[22] E. J. Candès and D. L. Donoho. Continuous curvelet transform: II. discretization and frames. *Appl. Comput. Harmon. Anal.*, 19:198–222, 2003.

[23] E. J. Candès and D. L. Donoho. Continuous curvelet transform: II. discretization and frames. *Appl. Comput. Harmon. Anal.*, 19:198–222, 2003.

[24] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise-C^2 singularities. *Comm. on Pure and Appl. Math*, 57:219–266, 2004.

[25] E. J. Candès and D. L. Donoho. Fast Discrete Curvelet Transfoms. *Multiscale Modeling and Simulation*, 5(3):861–899, 2006.

[26] H.T. Chen, T.L. Liu, and C.S. Fuh. Segmenting highly articulated video objects with weak-prior random forests. *Lecture Notes in Computer Science: ECCV 2006 proceedings*, 3954:373–385, 2006.

[27] S.C. Chen, G. J. Gordon, and R.F. Murphy. Graphical models for structured classification, with an application to interpreting images of protein subcellular location patterns. *Journal of Machine Learning Research*, 9:651–682, 2008.

[28] Wen-Chang Cheng and Ding-Mao Jhan. A self-constructing cascade classifier with adaboost and svm for pedestrian detection. *Engineering Applications of Artificial Intelligence*, 26:1016–1028, 2013.

[29] C.K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, 6:247–254, 1957.

[30] C.K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[31] Echeverri CJ. and Perrimon N. High-throughput rnai screening in cultured cells: a user's guide. *Nature Review Genetics*, 7:373–384, 2006.

[32] David A. Clausi. An analysis of co-occurence texture statistics as a function of grey level quantization. *Can. J. Remote Sensing*, 28(1):45–62, 2002.

[33] Bishop C.M. *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.

[34] Andre L.V. Coelho and Diego S.C. Nascimento. On the evolutionary design of heterogeneous bagging models. *Neuralcomputing*, pages 3319–3312, 2010.

[35] Filipe Condessa, José Bioucas-Dias, Carlos A. Castro, John A. Ozolek, and Jelena Kovačevíc. Classification with reject option using contextual information. In *Proceedings of 2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 1340–1343, 2013.

[36] Christophe Croux, Kristel Joossens, and Aurelie Lemmens. Trimmed bagging. *Computational Statistics & Data Analysis*, 52(1):362–368, 2007.

[37] Zhang C.X. and Zhang J.S. A local boosting algorithm for solving classification problems. *Computational Statistics & Data Analysis*, 52:1928–1941, 2008.

[38] Margineantu D. and Dietterich T. Pruning adaptive boosting. In *Proceedings of 14th International Conference on Machine Learning*, pages 211–218, 1997.

[39] Opitz D. and Shavlik J. Generating accurate and diverse members of neural network ensemble. In *Avances in neural information processing systems vol. 8*, pages 535–541. The MIT Press, Cambridge, 1996.

[40] Opitz D. and Maclin R. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, pages 169–198, 1999.

[41] Jean-Romain Dalle, Wee Kheng Leow, Daniel Racoceanu, Adina Eunice Tutac, and Thomas C. Putti. Automatic breast cancer grading of histopathological images. In *Proc. 30th Annual international IEEE EMBS conference*, pages 3052–3055. IEEE, 2008.

[42] I. Daubechies. *Ten Lectures on Wavelets*, page 137. SIAM, Philadelphia,PA, 1992.

[43] T. G. Dietterich. Ensemble methods in machine learning. *in MCS'00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[44] T.G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[45] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski. Automated grading of prostate cancer using architectural and textural image features. In *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1284–1287, 2007.

[46] Scott Doyle, Michael Feldman, John Tomaszewski, and Anant Madabuhushi. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Transactions on Biomedical Engineering*, 59(5):1205–1218, 2012.

[47] Scott Doyle, Michael Feldman, John Tomaszewski, Natalie Shih, and Anant Madabhushi. Cascaded multi-class pairwise classifier (cascampa) for normal, cancerous, and cancer confounder classes in prostate histology. In *Proceedings of the ISBI 2011*, pages 715–718, 2011.

[48] Scott Doyle, Michael D Feldman, Natalie Shih, John Tomaszewki, and Anant Madabhushi. Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. *BMC Bioinformatics*, 13(282):1–15, 2012.

[49] M. Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N. Gurcan. Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology. *IEEE Transactions on Biomedical Engineering*, 58(7):1977–1984, 2011.

[50] Alpaydin E. and Kaynak C. Cascading classifiers. *Kybernetika*, 34:369–374, 1998.

[51] Schapire R. E. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[52] Colm Elliott, Douglas L. Arnold, D. Louis Collins, and Tal Arbel. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain mri. *IEEE Transactions on Medical Imaging*, 32(8):1490–1503, 2013.

[53] T. Evgeniou, M. Pontil, and A. Elisseeff. Leave one out error, stability, and generalization of voting combinations of classifiers. *Machine Learning*, 55:71–97, 2004.

[54] P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Multiclassification: reject criteria for the baysian combiner. *Pattern Recognition*, 32:1435–1447, 1999.

[55] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, 28 (2):337–407, 2000.

[56] G. Fumera. *The advanced methods for pattern recognition with the reject option.* PhD thesis, University of Caliary, Caliary, Italy, 2002.

[57] Giorgio Fumera and Fabio Roli. Support vector machines with embedded reject option. In *Proceedings of the Int. Workshop on Pattern Recognition with Support Vector Machines (SVM2002), Niagara Falls*, pages 68–82. Springer, 2002.

[58] Giorgio Fumera and Fabio Roli. Analysis of error-reject trade-off in linear combined multiple classifiers. *Pattern Recognition*, 37:1245–1265, 2004.

[59] Giorgio Fumera, Fabio Roli, and Giorgio Giacinto. Reject option with multiple thresholds. *Pattern Recogntion*, 33:2099–2101, 2000.

[60] Giorgio Fumera, Fabio Roli, and Alessandra Serrau. A theoretical analysis of bagging as a linear combination of classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1293–1299, 2008.

[61] Jorge García-Gutiérrez, Daniel Mateos-Garcí, and José C. Riquelme-Santos. Evor-stack: A label-dependent evolutive stacking on remote sensing data fusion. *Neurocomputing*, 75:115–122, 2012.

[62] G. Giacinto and F. Roli. Dynamic classifier selection based on multiple classifier behaviour. *Pattern Recognition*, 34(9):1879–1881, 2001.

[63] N. Giusti, F. Masulli, and A. Sperduti. A theoretical and experimental analysis of a two-stage system for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:893–904, 2002.

[64] Hannon G.J. Rna interference. *Nature*, 418:244–251, 2002.

[65] King-Shy Goh, Edward Y. Chang, and Beitao Li. Using one-class and two-class svms for multiclass image annotation. *IEEE Transactions on Knowledge and Data Engineering*, 17(10):1333–1346, 2005.

[66] Lena Gorelick, Olga Veksler, Mena Gaed, José A. Gómez, Madeleine Moussa, Glenn Bauman, Aaron Fenster, and Aaron D. Ward. Prostate histopathology: Learning tissue component histograms for cancer detection and classification. *IEEE Transactions on Medical Imaging*, 32(10):1804–1818, 2013.

[67] Zhenhua Guo, Lei Zhang, and David Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

[68] Metin N. Gurcan, Laura E. Boucheron, Ali Can, Anant Madabhushi, Nasir M. Rajpoot, and Bulent Yener. Histopathological Image Analysis: A Review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[69] Metin N. Gurcan, Laura E. Boucheron, Ali Can, Anant Madabhushi, Nasir M. Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, 2:147–171, 2009.

[70] Wolpert D. H. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[71] Zhou Z. H., Wu J., and Tang W. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137:239–263, 2002.

[72] Mehdi Salkhordeh Haghighi, Abedin Vahedian, and Hadi Sadoghi Yazdi. Creating and measuring diversity in multiple classifier systems using support vector data description. *Applied Soft Computing*, 11:4931–4942, 2011.

[73] P. Han, X. Zhang, R.S. Norton, and Z.P. Feng. Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinformatics*, 10(8):1–9, 2009.

[74] Blaise Hanczar and Edward R. Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–1895, 2010.

[75] H. Hao, C.-L. Liu, and H. Sako. Confidence evaluation for combining diverse classifiers. In *Proceedings of Seventh International Conference on Document Analysis and Recognition*, pages 760–764. IEEE Computer Society, 2003.

[76] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. *IEEE Trans. Systems, Man and Cybernetics*, 3(6):610–621, 1973.

[77] S. Haykin. *An Introduction to Neural Networks-A Comprehensive Foundation, 2nd Edition.* Prentice-Hall, Upper Saddle River, NJ, 1999.

[78] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[79] Heiko Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 40:863–874, 2007.

[80] H. Gholam Hosseini, K. J. Reynolds, and D. Powers. A multi-stage neural network classifier for ecg events. In *Proceedings of the 23rd Annual EMBS International Conference*, pages 1672–1675, 2001.

[81] Yanhua Hu, Jesus Carmona, and Robert F. Murphy. Application of temporal texture features to automated analysis of protein subcellualr locations in time serie fluorescene images. In *Proceedings of the ISBI 2006*, pages 1028–1031, 2006.

[82] Yanhua Hu, Elvira Osuna-Highley, Juchang Hua, Theodore Scott Nowicki, Robert Stolz, Camille McKayle, and Robert F. Murphy. Automated analysis of protein subcellular location in time series images. *Bioinformatics*, 26(13):1630–1636, 2010.

[83] Chao-Hui Huang, Antoine Veillard, Ludovic Roux, Nocolas Loménie, and Daniel Rcoceanu. Time-efficient sparse analysis of histopathological while slide images. Preprint (2010) **dio**:10.1016/j.compmedimag.2010.11.009, 2010.

[84] Wang J., Zhou X., Bradley P.L., Chang S.F., Perrimon N., and Wong S.T.C. Cellular phenotype recognition for high-content rna interference genome-wide screening. *Journal of Biomolecular Screening*, 13:29–39, 2008.

[85] Kourosh Jafari-Khouzani and Hamid Soltanian-Zadeh. Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering*, 50(6):697–704, 2003.

[86] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[87] Clemens JC., Worby CA., Simonson-Leff N., Muda M., Maehama T., Hemmings BA., and Dixon JE. Use of double-stranded rna interference in drosophila cell lines to dissect signal transduction pathways. *Proceedings of the Natural Academy of Sciences*, 97:6499–6503, 2000.

[88] Yarrow J.C., Feng Y., Perlman Z.E., Kirchhausen T., and Mitchison T.J. Phenotypic screening of small molecule libraries by high throughput cell imaging. *Combinatorial Chemistry & High Throughput Screening*, 6:279–286, 2003.

[89] Shi jin Wang, Avin Mathew, Yan Chen, Li feng Xi, Lin Ma, and Jay Lee. Empirical analysis of support vector machine ensemble classifiers. *Expert Systems and Applications*, 36:6466–6476, 2009.

[90] Chan P. K. and Stolfo S. J. A comparative evaluation of voting and meta-learning on partitioned data. In *Proceedings of the 12th International Conference on Machine Learning*, pages 90–98. Morgan Kaufumann, 1995.

[91] Chan P. K. and Stolfo S. J. On the accuracy of meta-learning for scalable data mining. *Journal of Intelligent Information Systems*, 8:5–28, 1997.

[92] Huang K. and Murphy RF. Automated classification of subcellular patterns in multicell images without segmentation into single cells. In *Proceedings of the ISBI 2004*, pages 1139–1142, 2004.

[93] Huang K. and Murphy RF. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics*, 5(78):1–19, 2004.

[94] Maya Kallas, Paul Honeine, Cédric Richard, Clovis Francis, and Hassan Amoud. Non-negativity constraints on the pre-image for pattern recognition with kernel machines. *Pattern Recognition*, 46:3066–3080, 2013.

[95] Cenk Kaynak and Ethem Alpaydin. Multistage cascading of multiple classifiers: one man's noise is another man's data. *Proceedings of ICML 2000*, pages 455–462, 2000.

[96] Shehroz S. Khan and Michael G. Madden. A survey of recent trends in one class classification. In Lorcan Coyle and Jill Freyne, editors, *Artificial Intelligence and Cognitive Science*, volume 6206 of *Lecture Notes in Computer Science*, pages 188–197. Springer Berlin Heidelberg, 2010.

[97] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S. Y. Bang. Pattern classification using support vector machine ensemble. *Proceedings of the 16th International Conference on Pattern Recognition*, 2:160–163, 2002.

[98] Young-Won Kim and Il-Seok Oh. Classifier ensemble selection using hybrid genetic algorithms. *Pattern Recognition Letters*, 29:796–802, 2008.

[99] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, and Jiri Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[100] Albert H.R. Ko, Robert Sabourin, and Alceu Souza Britto Jr. From dynamic classi?er selection to dynamic ensemble selection. *Pattern Recognition*, 41:1718–1731, 2008.

[101] Sonal Kothari, John H Phan, Andrew N Young, and May D Wang. Histological image classification using biological interpretable shape-based features. *BMC Medical Imaging*, 13(9):1–16, 2013.

[102] Bartosz Krawczyk. Diversity in ensembles for one-class classification. In Mykola Pechenizkiy and Marek Wojciechowski, editors, *Advances in Intelligent Systems and Computing*, volume 185 of *New Trends in Databases and Information Systems*, pages 119–129. Springer Berlin Heidelberg, 2013.

[103] M. Muthu Rama Krishnan, Vikram Venkatraghavan, U. Rajendra Acharya, Mousumi Pal, Ranjan Rashmi Paul, Lim Choo Min, Ajoy Kumar Ray, Jyotirmoy Chatterjee, and Chandan Chakraborty. Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm. *BMC Bioinformatics*, 13(282):1–15, 2012.

[104] L. Kuncheva and L. C. Jain. Designing classifier fusion systems by genetic algorithms. *IEEE Transactions on Evolutionary Computation*, 4(4):327–336, 2000.

[105] L. I. Kuncheva, J. Bezdek, and R. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34:299–314, 2001.

[106] L.I. Kuncheva and J.J. Rodríguez. Classifier ensembles with a random linear oracle. *IEEE Transactions on Knowledge and Data Engineering*, 19(4):500–508, 2007.

[107] L.I. Kuncheva, J.J. Rodriguez, C.O. Plumpton, D.E. Linden, and S.J. Johnston. Random subspace ensembles for fmri classification. *IEEE Transactions on Medical Imaging*, 29(2):531–542, 2010.

[108] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. NJ: Wiley, 2004.

[109] Marek W. Kurzynski. On the multistage bayes classifier. *Pattern Recognition*, 21(4):355–365, 1988.

[110] James Tin-Yau Kwok and Ivor Wai-Hung Tsang. The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, 15(6):1517–1525, 2004.

[111] Breiman L. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[112] Rokach L. Decomposition methodology for classification tasks - a meta decomposition framework. *Pattern Analysis and Applications*, 9:257–271, 2006.

[113] Shamir L., Orlov N., Eckley D.M., Macura T., and Goldberg I. Iicbu 2008 - a proposed benchmark suite for biological image analysis. *Medical & Biological Engineering & Computing*, 46:943–947, 2008.

[114] L. Lam and C.Y. Suen. Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics -Part A: Systems and Human*, 27:553–568, 1997.

[115] Peng Li, Kap Luk Chan, Sheng Fu, and Shankar M. Krishnan. An abnormal ecg beat detector approach for long-term monitoring of heart patients based on hybrid kernel machine ensemble. In *Proc. International Workshop on Multiple Classifier Systems (MCS 2005)*, pages 346–355. Springer Verlag, 2005.

[116] Peng Li, Kap Luk Chan, and Shankar M. Krishnan. Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection in colonoscopic images. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, pages 670–675. IEEE Computer Society, 2005.

[117] Xirong Li, Snoek C.G.M., Worring M., Koelma D., and Smeulders A.W.M. Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia*, 15 (4):933–945, 2013.

[118] Eleni Linos, Demetri Spanos, Bernard A. Rosner, Katerina Linos, Therese Hesketh, Jian Ding Qu, Yu-Tang Gao, Wei Zheng, and Graham A. Colditz. Effects of Reproductive and Demographic Changes on Breast Cancer Incidence in China: A Modeling Analysis. *Journal of the National Cancer Institute*, 100:1352–1360, 2008.

[119] Suzanne Little, Sara Colantonio, Ovidio Salvetti, and Petra Perner. Evaluation of feature subset selection, feature weighting, and prototype selection for biomedical applications. *Journal of Software Engineering and Applications*, 3:39–49, 2010.

[120] Manhua Liu, Daoqiang Zhang, and Dinggang Shen. Ensemble sparse classification of alzheimer's disease. *NeuroImage*, 60(2):1106–1116, 2012.

[121] C. Loukas. A survey on histological image analysis-based assessment of three major biological factors influencing radiotherapy: proliferation, hypoxia and vascluature. *Computer Methods and Programs in Biomedicine*, 74(3):183–199, 2004.

[122] Cordella L.P., C. De Setfano, F. Tortorella, and M. Vento. A method for improving classification reliability of multilayer perceptrons. *IEEE Transactions on Neural Networks*, 6:1140–1147, 1995.

[123] Huitao Luo. Optimization design of cascaded classifiers. In *Proceedings of the CVPR 2005*, pages 480–485, 2005.

[124] Boutros M., Kiger A.A., Armknecht S., Kerr K., Hild M., Koch B., Haas S.A., Paro R., and Perrimon N. Genome-wide rnai analysis of growth and viability in drosophila cells. *Science*, 303:832–835, 2004.

[125] Islam M. M., Yao X., and Murase K. A constructive algorithm for training cooperative neural network ensembles. *IEEE Transactions on Neural Networks*, 14(4):820–834, 2003.

[126] Skurichina M. and Duin R. P. W. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis and Applications*, 5(2):121–135, 2002.

[127] Wezel M. and Potharst R. Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, 181:436–452, 2007.

[128] Jianwei Ma and Gerlind Plonka. The curvelet transform. *IEEE Signal Processing Magazine*, 27(2):118–133, 2010.

[129] Larry M. Manevitz and Malik Yousef. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154, 2001.

[130] Markos Markou and Sameer Singh. Novelty detection: a review-part 1: statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

[131] Markos Markou and Sameer Singh. Novelty detection: a review-part 2: neural network based approaches. *Signal Processing*, 83:2499–2521, 2003.

[132] Eitan Menahem, Lior Rokach, and Yuval Elovici. Troika c an improved stacking schema for classification tasks. *Information Science*, 179:4097–4122, 2009.

[133] S. Merler, B. Caprile, and C. Furlanello. Parallelizing adaboost by weights dynamics. *Computational Statistics & Data Analysis*, 51:2487–2498, 2007.

[134] Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 536–542. IMIT Press Cambridge, MA, USA, 1998.

[135] T. Mitchell. *Machine Learning*. New York: McGraw Hill, 1997.

[136] Anderson M.L and Oates T. review of recent research in metareasoning and metalearning. *AI Magazine*, 28:7–16, 2007.

[137] M. Moya, M. Koch, and L. Hostetler. One-class classifier networks for target recognition applications. In *Proceedings of World Congress on Neural Networks*, pages 797–801, 1993.

[138] Weirauch M.T., Wong C.K., Byrne A.B., and Stuart J.M. Information-based methods for predicting gene function from systematic gene knock-downs. *BMC Bioinformatics*, 9:1–21, 2008.

[139] Boland M.V., Markey M., and Murphy R.F. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 33:366–375, 1998.

[140] Boland M.V. and Murphy R.F. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. *Bioinformatics*, 17:1213–1223, 2001.

[141] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 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*, pages 65–81, 2010.

[142] Bingbing Ni, Shuicheng Yan, Meng Wang, Kassim A.A., and Qi Tian. High-order local spatial context modeling by spatialized random forest. *IEEE Transactions on Image Processing*, 22(2):739–751, 2013.

[143] Ilia Nouretdinov, Sergi G. Costafreda, Alexander Gammerman, Alexey Chervonenkis, Vladimir Vovk, Vladimir Vapnik, and Cynthia H.Y. Fu. Machine learning classification with confidence: Application of transductive conformal predictors to mri-based diagnostic and prognositc markers in depression. *NeuroImage*, 56:809–813, 2011.

[144] Maimon O. and Rokach L. Improving supervised learning by feature decomposition. In *Proc. of foundations of information and knowledge systems*, pages 178–196. Springer Verlag, 2002.

[145] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[146] O. Okun and H. Priisalu. Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors. *Artificial Intelligence in Medicine*, 45:151–162, 2009.

[147] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D. Mark Eckley, and Ilya G. Goldberg. Wnd-charm: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693, 2008.

[148] Buhlmann P. and Hothorn T. Twin boosting: Improve feature selection and prediction. *Statistics and Computing*, 20(2):119–138, 2010.

[149] Clark P. and Boswell R. Rule induction with cn2: some recent improvements. In *Proceedings of the European working session on learning*, pages 151–163. Pitman, 1991.

[150] Derbeko P., El-Yaniv R., and Meir R. Variance optimized bagging. In *Proceedings of European Conference on Machine Learning 2002*, pages 60–71. Springer Verlag, London, 2002.

[151] Pudil P., Novovicova J., Blaha S., and Kittler J. Multistage pattern recognition with reject option. *Proceeding of the Eleventh IAPR International Conference on Pattern Recognition B*, pages 92–95, 1995.

[152] Roberto Perdisci and Guofei Gu. Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems. In *Proceedings of the IEEE International Conference on Data Mining (ICDM06)*, pages 488–498. IEEE Computer Society, 2006.

[153] Hu Q., Yu D., Xie Z., and Li X. Eros: ensemble rough subspaces. *Pattern Recognition*, 40:3728–3739, 2007.

[154] Y. Qian and R.F. Murphy. Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models. *Bioinformatics*, 24:569–576, 2008.

[155] José R. Quevedo, Antonio Bahamonde, Miguel Pérez-Enciso, and Oscar Luaces. Disease liability prediction from large scale genotyping data using classifiers with a reject option. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):88–98, 2012.

[156] H. Qureshi, O. Sertel, N. Rajpoot, R. Wilson, and M. Gurcan. Adaptive discriminant wavelet package transform and local binary patterns for meningioma subtype classification. In *Proceedings of MICCAI 2008*, pages 196–204. Springer Verlag, 2008.

[157] Bryll R., Gutierrez-Osuna R., and Quek F. Bagging: improving accuracy of classifier ensembles by using random feature subset. *Pattern Recognition*, 36:1291–1302, 2003.

[158] Caruana R., Niculescu-Mizil A., Crew G., and Ksikes A. Ensemble selection from libraries of models. In *Proceedings of 21th International Conference on Machine Learning*, page 18, 2004.

[159] Nurettin Acır, İbrahim Öztura, Mehmet Kuntalp, Brış Baklan, and Cüneys Güzeliş. Automatic detection of epileptiform events in eeg by a three-stage procedure based on artificial neural networks. *IEEE Transactions on Biomedical Engineering*, 52(1):30–40, 2005.

[160] Quinlan J. R. *C4.5: programs for machine learning*. Morgan Kaufmann, Los Altos, 1993.

[161] J. Ramirez, J.M. Gorriz, R. Chaves, M. Lopez, D. Salas-Gonzalez, I. Alvarez, and F. Segovia. Spect image classification using random forests. *Electronics Letters*, 45(12):604–605, 2009.

[162] De Ridder, D.M.J. Tax, and D. Duin. An experimental comparison of one-class classification methods. In *Proceedings of the 4th Annual Conference of the Advanced School for Computing and Imaging, Delft*, pages 213–218, 1998.

[163] Irina Rish. An empirical study of the naive bayes classifier. In *Proceedings of the IJCAI 2001*, pages 41–46, 2001.

[164] J.J. Rodriguez, L.I. Kuncheva, and Carlos J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.

[165] R.O.Duda, P.E.Hart, and D.G.Stork. *Pattern Classification, 2nd ed.* New York: Wiley, 2001.

[166] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33:1–19, 2010.

[167] Niall Rooney and David Patterson. A weighted combination of stacking and dynamic integration. *Pattern Recognition*, 40:1385–1388, 2007.

[168] David Rotger, Petia Radeva, and Nico Bruining. Automatic detection of bioabsorbable coronary stents in ivus images using a cascade of classifiers. *IEEE Transactions on Information Techonology in Biomedicine*, 14(2):535–537, 2010.

[169] Volker Roth. Kernel fisher discriminants for outlier detection. *Neural Computation*, 18:942–960, 2006.

[170] Džeroski S. and Ženko B. Is combining classifiers with stack better than selecting the best one? *Machine Learning*, 54(3):255–273, 2004.

[171] Mohammad Javad Saberian and Nuno Vasconcelos. Learning optimal embedded cascades. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2005–2018, 2012.

[172] Andrea Sboner, Claudio Eccher, Enrico Blanzieri, Paolo Bauer, Mario Cristofolini, Giuseppe Zumiani, and Stefano Forti. A multiple classifier system for early melanoma diagnosis. *Artificial Intelligence in Medicine*, 27(1):29–44, 2003.

[173] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidencerated predictions. *Machine Learning*, 37(3):297–336, 1999.

[174] Bernhard Schölkopf. The kernel trick for distances. Technical Report MSR-TR-2000-51, Microsoft Research, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, 2000.

[175] Bernhard Schölkopf, J. Platt, John Shawe-Taylor, Alex Smola, and Robert C. Williamson. Estimating the support of a high dimensional distribution. *Neural Computation*, 13(7):1443–1472, 2001.

[176] Alexander K. Seewald and Johannes Fürnkranz. An evaluation of grading classifiers. In *Proceedings of 4th International Conference on Advanced Data Analysis*, pages 115–125, 2001.

[177] A.J.C. Sharky. *Combining artificial neural nets. Ensemble and modular multi-net systems.* Springer Verlag, London, 1999.

[178] Albert D. Shieh and David F. Kamm. Ensembles of one class support vector machines. In *Proc. Multiple Classifier Systems, 2009*, pages 181–190. Springer Verlag, 2009.

[179] P. Simeone, C. Marrocco, and F. Tortorella. Design of reject rules for ecoc classification systems. *Pattern Recognition*, 45:863–875, 2012.

[180] Marina Skurichina and Robert P. W. Duin. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. *Pattern Analysis & Applications*, 5:121–135, 2002.

[181] American Cancer Society. *Breast Cancer Facts & Figures 2011-2012*. American Cancer Society, Inc., 2011.

[182] Leen-Kiat Soh and Costas Tsatsoulis. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. *IEEE Trans. Geoscience and Remote Sensing*, 37(2):780–795, 1999.

[183] Claudio De Stefano, Carlo Sansone, and Mario Vento. To reject or not reject: That is the question - an answer in case of neural networks. *IEEE Transactions on Sytem, Man and Cybernetics-Part C: Applications and Reviews*, 30(1):84–94, 2000.

[184] Ishrat Jahan Sumana, Md. Monirul Islam, Dengsheng Zhang, and Guojun Lu. Content based image retrieval using curvelet transform. In *IEEE 10th workshop on Multimedia signal processing*, pages 11–16, 2008.

[185] Zhenan Sun, Yunhong Wang, Tieniu Tan, and Jiali Cui. Improving iris recognition accuracy via cascaded classifiers. *IEEE Transactions on Systems, Man and Cybernetics–Part C: Applications and Reviews*, 35(3):435–441, 2005.

[186] Jolliffe I. T. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[187] Ali Tabesh, Mikhail Teverovskiy, Ho-Yuen Pang, Vinay P. Kumar, David Verbel, Angeliki Kotsianti, and Olivier Saidi. Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Transactions on Medical Imaging*, 26(10):1366–1378, 2007.

[188] Yuchun Tang, Yan-Qing Zhang, and Zhen Huang. Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3):365–381, 2007.

[189] David M.J. Tax and Robert P.W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20:1191–1199, 1999.

[190] David M.J. Tax and Robert P.W. Duin. Combining one-class classifiers. In *Proc. Multiple Classifier Systems, 2001*, pages 299–308. Springer Verlag, 2001.

[191] David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004.

[192] D.M.J. Tax. *One-Class Classification*. PhD thesis, Delft University of Technology, 2001.

[193] D.M.J. Tax and R.P.W. Duin. Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29:1565–1570, 2008.

[194] Hong Tian, Zhu Duan, Ajith Abraham, and Hongbo Liu. A novel multiplex cascade classifier for pedestrian detection. *Pattern Recognition Letters*, 34:1687–1693, 2013.

[195] Ha T.M. The optimum class-selective rejection rule. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:608–615, 1997.

[196] G. Tsoumakas, I. Partalas, and I. Vlahavas. A taxonomy and short review of ensemble selection. In *ECAI2008, workshop on Supervised and Unsupervised Ensemble Methods and Their Applications*, pages 41–46, 2008.

[197] Te-Ming Tu, Chin-Hsing Chen, Jiunn-Lin Wu, and Chein-I Chang. A fast two-stage classification method for high-dimensional remote sensing data. *IEEE Transactions on Gioscience and remote sensing*, 36(1):182–192, 1998.

[198] K. Tumer and N. C. Oza. Input decimated ensembles. *Pattern Analysis and Applications*, 6(1):65–77, 2003.

[199] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of CVPR 2001*, pages 511–518, 2001.

[200] Vapnik V.N. *The Nature of Statistical Learning Theory (2nd Ed.)*. Sringer Verlag, 2000.

[201] Tukey J. W. *Exploratory data analysis*. Addison-Wisley, Reading, 1977.

[202] Q. Wang, L.S. Lopes, and D.M.J. Tax. Visual object recognition through one-class learning. In *International Conference on Image Analysis and Recognition*, pages 463–470, 2004.

[203] Marten H. Wegkamp and Ming Yuan. Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385, 2011.

[204] Liyang Wei, Yongyi Yang, Robert M. Nishikawa, Miles N. Wernick, and Alexandra Edwards. Relevance vector machine for automatic detection of clustered microcalcifications. *IEEE Transactions on Medical Imaging*, 24(10):1278–1285, 2005.

[205] Jia-Bao Wen, Yue-ShanXiong, and Shu-LinWang. A novel two-stage weak classifier selection approach for adaptive boosting for cascade face detector. *Neurocomputing*, 116:122–135, 2013.

[206] C.K.I. Williams. On a connection between kernel pca and metric multidimensional scaling. In *Advances in Neural Information Processing Systems 13*, NIPS 2001, pages 675–681. MIT Press, 2001.

[207] K. Woods, P. Kegelmeyer, and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (4):405–410, 1997.

[208] Jin Xiao, Changzheng He, Xiaoyi Jiang, and Dunhu Liu. A dynamic classifier ensemble selection approach for noise data. *Information Science*, 180:3402–3421, 2010.

[209] Tong Xiao, Jingbo Zhu, and Tongran Liu. Bagging and boosting statistical machine translation systems. *Artificial Intelligence*, 195:496–527, 2013.

[210] Li Xiaohua, Kin-Man Lam, Shen Lansun, and Zhou Jiliu. Face detection using simplified gabor features and hierarchical regions in a cascade of classifiers. *Pattern Recognition Letters*, 30:717–728, 2009.

[211] Freund Y. and Schapire R. E. Experiments with a new boosting algorithm. In *Proceedings of the 13th international conference on machine learning*, pages 325–332, 1996.

[212] Le Cun Y., B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Handwritten digit recognition with a back propagation network. In *Advances in Neural Information Processing Systems II*, pages 595–601, San Mateo, 1990. Morgan Kaufuman.

[213] Pengyi Yang, Yee Hwa Yang, Bing B. Zhou, and Albert Y. Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010.

[214] Gang Yu, N.A. Goussies, Junsong Yuan, and Zicheng Liu. Fast action detection via discriminative random forest voting and top-k subvolume search. *IEEE Transactions on Multimedia*, 13(3):507–517, 2011.

[215] Jun Yu, Feng Lin, Hock-Soon Seah, Cuihua Li, and Ziyu Lin. Image classification by multimodal subspace learning. *Pattern Recognition Letters*, 33:1196–1204, 2012.

[216] Perlman Z.E., Slack M.D., Feng Y., Mitchison T.J., Wu L.F., and Altschule S.J. Multidimensional drug profiling by automated microscopy. *Science*, 306:1194–1198, 2004.

[217] Zhou Z.H. and Tang W. Selective ensemble of decision trees. In *Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, pages 476–483. Springer Verlag, 2003.

[218] Bailing Zhang, Yungang Zhang, Wenjin Lu, and Guoxia Han. Phenotype Recognition by Curvelet Transform and Random Subspace Ensemble. *Journal of Applied Mathematics and Bioinformatics*, 1:79–103, 2011.

[219] Lefei Zhang, Liangpei Zhang, Dacheng Tao, and Xin Huang. On combining multiple features for hyperspectral remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3):879–893, March 2012.

[220] Ping Zhang, Tien D. Bui, and Ching Y. Suen. A novel cascade ensemble classifier system with a high recognition performance on handwritten digits. *Pattern Recognition*, 40:3415–3429, 2007.

[221] Ping Zhang and Xi Guo. A cascade face recognition system using hybrid feature extraction. *Digital Signal Processing*, 22:987–993, 2012.

[222] Rong Zhang and Dimitris N. Metaxas. Ro-svm: Support vector machine with reject option for image categorization. In *BMVC'06*, pages 1209–1218, 2006.

[223] Yungang Zhang, Bailing Zhang, Frans Coenen, and Wenjin Lu. Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles. *Machine Vision and Applications*, DOI 10.1007/s00138-012-0459-8:1–17, 2012.

[224] Yungang Zhang, Bailing Zhang, Frans Coenen, and Wenjin Lu. Highly reliable breast cancer diagnosis with cascaded ensemble classifiers. In *Proceedings of the 2012 International Joint Conference on Neural Networks*, pages 1–8. IEEE Computational Intelligence Society, 2012.

[225] Yungang Zhang, Bailing Zhang, Frans Coenen, and Wenjin Lu. Cascading One-Class Kernel Subspace Ensembles for Reliable Biopsy Image Classification. *Journal of Medical Imaging and Health Informatics*, 4:1–12, 2014.

[226] Yungang Zhang, Bailing Zhang, Frans Coenen, and Wenjin Lu. One-Class Kernel Subspace Classifier Ensemble for Medical Image Classification. *Eurasip Journal on Advances in Signal Processing*, 2014:16:1–13, 2014.

[227] Yungang Zhang, Bailing Zhang, and Wenjin Lu. Breast cancer classification from histological images with multiple features and random subspace classifier ensemble. In *Proceedings of the 2011 conference on Computational Models for Life Sciences*, pages 19–28. American Institute of Physics, 2011.

[228] Yungang Zhang, Bailing Zhang, and Wenjin Lu. Breast Cancer Histological Image Classification with Multiple Features and Random Subspace Classifier Ensemble. *T.D. Pham, L.C. Jain (eds): Innovations in Knowledge-based Systems in Biomedicine*, SCI 450:27–42, 2013.

[229] Wei-Shi Zheng, JianHuang Lai, and Pong C. Yuen. Penalized preimage learning in kernel principle component analysis. *IEEE Transactions on Neural Networks*, 21(4):551–570, 2010.

[230] Changfang Zhu, Elizabeth S. Burnside, Gale A. Sisney, Lonie R. Salkowski, Josephine M. Harter, Bing yu, and Nirmala Ramanujam. Fluorescence spectroscopy: An ajunct diagnostic tool to image-guided core needle biopsy of the breast. *IEEE Transactions on Biomedical Engineering*, 56(10):2518–2528, 2009.