

Phenotype Recognition by Curvelet Transform and Random Subspace Ensemble

Bailing Zhang¹, Yungang Zhang², Wenjin Lu³ and Guoxia Han⁴

Abstract

Automated, image based high-content screening has become a fundamental tool for scientists to make discovery in biological science. Modern robotic fluorescence microscopes are able to capture thousands of images from massively parallel experiments such as RNA interference (RNAi). As such, efficient methods are required for automatic cellular phenotype identification capable of dealing with large image data sets. In this paper we applied the Curvelet transform for image feature description and Random Subspace ensemble (RSE) for classification. The Curvelet transform as a new multiscale directional transform allows an almost optimal nonadaptive sparse representation of objects rich with edges. The RSE contains a set of base classifiers trained using randomly drawn subsets of curvelet features. The component classifiers are then aggregated by the Majority Voting Rule. Experimental results on the phenotype recognition from three benchmark fluorescence

¹ Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China, email: bailing.zhang@xjtlu.edu.cn

² Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China, email: yungang.zhang@xjtlu.edu.cn

³ Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China, e-mail: wenjin.lu@xjtlu.edu.cn

⁴ Department of Biological Science, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China, e-mail: guoxia.han@xjtlu.edu.cn

microscopy image sets (RNAi, CHO and 2D HeLa) show the effectiveness of the proposed approach. The ensemble model produces better performance compared to any of individual neural networks. It offers the classification rate 86.5% on the RNAi dataset, which compares favorably with the published result 82%, and the results on the other two sets of fluorescence microscopy images also confirm the effectiveness of the proposed approach.

Mathematics Subject Classification : 92C55, 68U10

Keywords: Phenotype Recognition, RNAi Screening, Curvelet Transform, Random subspace ensemble

1 Introduction

With appropriate staining techniques, complex cellular structures such as organelles within the eukaryotic cell can be studied by fluorescence microscopy images of cells. 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 [1]. 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 [2]. 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 [3,4].

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 [5]. 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 [6,7]. 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 [8]. 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.

To advance the development of high content screening for genome analysis, computer vision and pattern analysis techniques have to be resorted to characterize the morphological phenotypes quantitatively in order to identify genes and their dynamic relationships on a genome-wide scale [9-11]. Such a bioimage informatics framework would consist of several components: cellular segmentation, cellular morphology and texture feature extraction, cellular phenotype classification, and clustering analysis [1]. 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. Most of the proposed approaches for image-based high-content screening employed feature set which consist of different combinations of morphological, edge, texture, geometric, moment and wavelet features [12-18]. In recent years, computer vision has seen much progresses in various feature descriptions based on extracting and pooling local structural information from images, many of which have become "off-the-shelf" standard methods applicable to bioimages analysis.

In this article, our effort is made toward the recognition of phenotypes for high-content RNAi screening by using a benchmark fluorescence microscopy

images [12]. We will first show that the Curvelet Transform (CT) [19] as the latest research achievements on multiresolution analysis for image, is extremely efficient in feature representation for cellular images in RNAi screening. Multiresolution ideas have been extensively applied in various areas, and image analysis in particular. The most popular multiresolution analysis tool is the Wavelet Transform. By wavelet analysis, an image can be decomposed at different scales and orientations using a wavelet basis vector. The successes of wavelets has enkindled scientists' interests in further research of multiresolution and harmonic analysis, with a number of new multiresolution, multidimensional tools, such as contourlet and curvelet, developed in the past few years [19-21]. Curvelet transform can accurately capture edge information by taking the form of basis elements which exhibit very high directional sensitivity and are highly anisotropic. It has been shown that curvelet is well suited for representing images and the efficiency has been demonstrated in many tasks such as edge detection and denoising [22]. Recently, promising results have also been reported for face recognition and image retrieval [23-26].

Our main contribution for high-content RNAi screening is the proposal of an efficient phenotype classification method based on the curvelet features. In the past, many machine learning methods such as artificial neural networks and Support Vector Machine (SVM) have been utilized for the classification of subcellular protein location patterns in conjunction with various feature representation methods from fluorescence microscope images [14,13,15,16,18]. Multi-class phenotype images, however, are often featured with large intra-class variations and small inter-class similarities, which poses serious problems for simultaneous multi-class separation using the standard classification algorithms. And other rate-limiting factors challenging classifier design for multi-class phenotype images include the "small sample size" problem, which means that the number of available training samples per class is usually much smaller than the dimensionality of the feature space. In order to facilitate the task of classifier design, the dimensionality of the feature space is often reduced by means of a projection technique such as Principal Component Analysis (PCA) or Fisher's Linear Discriminant Analysis (LDA). The successes of such projection methods, however, are limited due to their inherent disadvantages.

To design accurate and robust classification systems for different applications, many advanced learning paradigms have been proposed in recent years.

Among them, the classification ensemble learning techniques have attracted much attention due to their demonstrated powerful capacities to improve upon the classification accuracy of a base learning algorithm. An ensemble of classifiers can integrate multiple component classifiers such as decision tree or multiple layer perceptrons for a same task using a same base learning algorithm [27,28,31]. To satisfy the diversity pre-requisite for the success of an ensemble system, namely, individual classifiers in an ensemble system making different errors on different instances, a common approach is to use different training datasets to train component classifiers. Such datasets are often obtained through re-sampling techniques, such as bootstrapping. To classify a new case, each member of the ensemble classifies the case independently of the others and then the resulting votes are aggregated to derive a single final classification. The representatives of ensemble learning include Bagging [29] and AdaBoost [30]. Diversity of classifiers in bagging is obtained by using bootstrapped replicas of the training data. That is, different training data subsets are randomly drawn with replacement from the entire training dataset. In AdaBoost, bootstrap training samples are drawn from a distribution that is iteratively updated such that subsequent classifiers focus on increasingly difficult instances. A version of bagging, called Random Forest [34], where the base classifier is a modified decision tree, has been proved to be efficient in many applications.

One popular classifiers ensemble method is the Random Subspace Ensemble (RSE), which was proposed in [35] based on a simple rule: constructing a different feature subset for each classifier in the ensemble. The main idea is to enhance the diversity among the component classifiers while keeping their accuracies at the same time. By using random feature subsets, RSE achieves some advantages for constructing and aggregating classifiers, especially in the case of that the number of available training objects is much smaller than the feature dimensionality. RSE is able to solve this small sample size problem. From another point of view, the RSE can avoid the dilemma of the curse of dimensionality. The dimensionality of each subspace is smaller than the original feature space dimensionality, and the number of training objects does not change. Therefore, the size of relative training samples increases. When data have redundant features, one may obtain better classifiers in random subspaces than in the original feature space. The combined decision of such

classifiers may be superior to a single classifier constructed on the original training set in the complete feature space. It has been demonstrated that it performs much better than several other ensemble methods such as bagging and boosting on many benchmark classification data sets [36]. Applications of Random Subspace Ensemble to medical and molecular image classification can be found in [37,38] and some improvements and variations have been discussed in [39,40]. With the RNAi, CHO and HeLa fluorescence microscopy image datasets [12], the improved classification performance will be demonstrated experimentally.

The remainder of this paper is organized as follows. Section 2 firstly introduces RNAi, CHO and 2D HeLa fluorescence microscopy image data which will be used and then details the Curvelet Transform for feature representation as well as the Random Subspace Ensemble algorithm. The experimental studies are presented in Section 3, and Section 4 offers the conclusions.

2 Fluorescence microscopy image data

Three benchmark fluorescence microscopy image datasets in [12] 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. According to [12], 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 Figure 1.

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

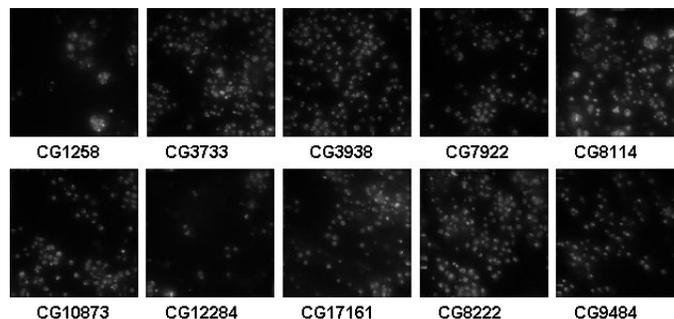


Figure 1: RNAi image set of fluorescence microscopy images of fly cells (*D. melanogaster*)

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 .

3 Curvelet Transform for Image Feature Description

Curvelet transform [19-21] is one of the latest developments of non-adaptive transforms today. Compared to wavelet, curvelet provides a more sparse representation of the image, with improved directional elements and better ability to represent edges and other singularities along curves. Compared to methods based on orthonormal transforms or direct time domain processing, sparse representation usually offers better performance with its capacity for efficient signal modelling. So far, successful applications of curvelet have been found in the fields of edge detection and image denoising [22], image retrieval [23,26] and face recognition [24]. The potential of curvelet transform to solve bioimage classification recognition problems, however, has not yet been explored.

While wavelets generalize the Fourier transform by using a basis that represents both location and spatial frequency, the curvelet transform provides the flexibility that the degree of localisation in orientation varies with scale. In curvelet, fine-scale basis functions are long ridges; the shape of the basis

functions at scale j is 2^{-j} by $2^{-j/2}$ so the fine-scale bases are skinny ridges with a precisely determined orientation. The curvelet coefficients can be expressed by

$$c(j, l, k) := \langle f, \varphi_{j,l,k} \rangle = \int_{\mathbb{R}^2} f(x) \varphi_{j,l,k}(x) dx \quad (1)$$

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

In the last few years, several discrete curvelet and curvelet-like transforms have been proposed. The influential approach is based on the Fast Fourier Transform (FFT) [21]. In the frequency domain, the curvelet transform can be implemented with φ by means of the window function U . Defining a pair of windows $W(r)$ (a radial window) and $V(t)$ (an angular window) as below:

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

$$\sum_{j=-\infty}^{\infty} V^2(t - 1) = 1, \quad t \in (-1/2, 1/2) \quad (3)$$

where variables W as a frequency domain variable, and r and θ as polar coordinates in the frequency domain, then for each $j \geq j_0$, U_j is defined in the Fourier domain by

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

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

The above brief introduction of the frequency plane partitioning into radial and angular divisions can be explained by Figure 2, where the shaded area is one of the polar wedges represented by U_j , which is supported by the radial and angular windows W and V . The radial divisions (concentric circles) are responsible for decomposition of the image in multiple scales (used for band-passing the image) and angular divisions corresponding to different angles or orientation. Therefore when we consider each wedge like the shaded one, one needs to define the scale and angle to analyze the bandpassed image at scale j and angle θ .

The technical details of the curvelet transform implementation is much involved and beyond the scope of current paper. The fastest curvelet transform currently available is vurvelets via wrapping [21,22], which will be used for our

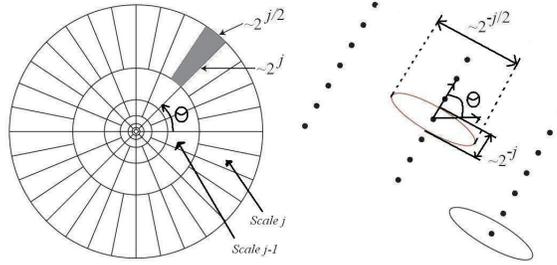


Figure 2: Curvelet transform: Fourier frequency domain partitioning (left) and spatial domain representation of a wedge (right)

work. If $f([t_1, t_2])$, $0 \leq t_1, t_2 \leq n$ is taken to be a Cartesian array and $\hat{f}[n_1, n_2]$ to denote its 2D Discrete Fourier Transform, then the architecture of curvelets via wrapping is as follows [20]:

1. 2D FFT (Fast Fourier Transform) is applied to obtain Fourier samples $\hat{f}[n_1, n_2]$, $-n/2 \leq n_1, n_2 < n/2$.
2. For each scale j and angle l , the product $\tilde{U}_{j,t}[n_1, n_2]\hat{f}[n_1, n_2]$ is formed, where $\tilde{U}_{j,t}[n_1, n_2]$ is the discrete localizing window.
3. Wrap the product around the origin to obtain $\tilde{f}[n_1, n_2] = W(\tilde{U}_{j,l}\hat{f})[n_1, n_2]$ where the range for n_1 and n_2 is now $0 \leq n_1 < L_{1,j}$, and $0 \leq n_2 < L_{2,j}$; $L_{1,j} \sim 2^j$ and $L_{2,j} \sim 2^{j/2}$ are constants.
4. Apply inverse 2D FFT to each $\tilde{f}_{j,l}$ thus creating the discrete curvelet coefficients

From the curvelet coefficients, some statistics can be calculated from each of these curvelet sub-bands as image descriptor. Similar to the Gabor filtering, the mean μ and standard deviation δ are the convenient features [26]. If n curvelets are used for the transform, $2n$ features $G = [G_\mu, G_\delta]$ are obtained, where $G_\mu = [\mu_1, \mu_2, \dots, \mu_n]$, $G_\delta = [\delta_1, \delta_2, \dots, \delta_n]$. The $2n$ dimension feature vector can be used to represent each image in the dataset.

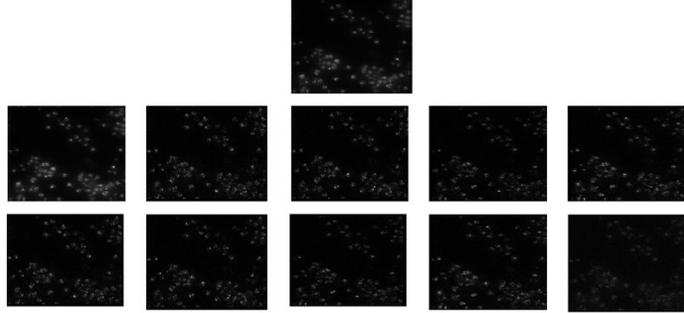


Figure 3: Curvelet transform of an RNAi microscopy image. The top image is the original one. The first image in second row is the approximate coefficients and others are detailed coefficients at eight angles from three scales. All the images are rescaled to same dimension for demonstration purpose.

4 Random Subspace Ensemble of Neural Network Classifiers

The idea of classifiers ensemble is to individually train a set of classifiers and appropriately combine their component decisions [27,28]. 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) [29] and re-weighting (boosting) [30] 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 component 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 former classifiers will play more important roles in the training of later classifiers.

Despite different classifier solutions can be applied in ensemble learning, in this paper we will consider only the 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 [32]. 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 existing of such differences between base classifiers is pre-requisite for the success of a classifier ensemble [28].

The multilayer perceptron (MLP) trained with the back propagation algorithm has been successfully applied to many classification problems in bioinformatics, for example, subcellular protein location patterns [13-15]. 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 using the back-propagation algorithm [32] to optimize a derivable criterion, such as the Mean Squared Error, using the available training data.

The performance improvement can be expected from an MLP ensemble by taking advantages of the disagreement among a set of MLP classifiers. An important issue in constructing the MLP ensemble is to create the diversity of the ensemble. In this paper we focus on MLP ensembles based on the Random Subspace, a successful ensemble generation technique [36]. Similar to random forest, the random subspace builds base classifiers independently with decision trees, each classifier is a decision tree, which are trained in a random reduced feature space [35]. The main idea of Random Subspace is: for a d -dimensional training set, choose a fixed n ($n < d$), randomly select n features according to the uniform distribution. Thus, the data of the original d -dimensional training set is transformed to the selected n -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 using majority voting. The main idea of Random Subspace is to simultaneously encourage diversity and individual accuracy within the ensemble: random fea-

ture 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 described as follows.

Consider a training set $X = \{X_1, X_2, \dots, X_n\}$, each training sample X_i is described by a p -dimensional vector, $X_i = \{x_{i1}, x_{i2}, \dots, x_{ip}\} (i = 1, \dots, n)$. 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 $X^r = \{X_1^r, X_2^r, \dots, X_n^r\}$, each training sample in X^r is described by a p^* feature vector, $X_i^r = \{x_{i1}^r, x_{i2}^r, \dots, x_{ip^*}^r\} (i = 1, \dots, n)$, where each feature component $x_{ij}^r (j = 1, \dots, p^*)$ 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 follows:

1. Repeat for $r = 1, 2, \dots, R$.
 - (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 .
 - (c) Compute combining weights $c_r = \frac{1}{2} \log \left(\frac{1 - err_r}{err_r} \right)$, where $err_r = \frac{1}{n} \sum_{i=1}^n w_i^r \xi_i^r$ and $\xi_i^r = 0$, if X_i is classified correctly. Otherwise, $\xi_i^r = 1$.
2. Combine classifiers $C^r(x), r = 1, \dots, R$, by the weighted majority vote with weights c_r to a final decision rule $\beta(x) = \operatorname{argmax}_{y \in \{-1, 1\}} \sum_r \delta_{\operatorname{sgn}(C^r(x)), y}$, where $\delta_{i,j} = 1$, if $i = j$. Otherwise, $\delta_{i,j} = 0$. $y \in \{-1, 1\}$ is a decision (class label) of the classifier.

5 Experiments

For the curvelet feature extraction process, fast discrete curvelet transform via wedge wrapping was applied to each of the images in the database using the CurveLab Toolbox (<http://www.curvelet.org/>), following the four steps described in Section 3: 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 5 levels analysis, $82 (= 1 + 16 + 32 + 32 + 1)$ subbands 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 (k NN), multi-layer perceptron neural networks, SVM and random subspace ensemble. k NN 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 now a standard multi-class classification method [32], which, in our comparisons, 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) [33] is a developed learning system originated from the statistical learning theory. 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 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 [34] 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

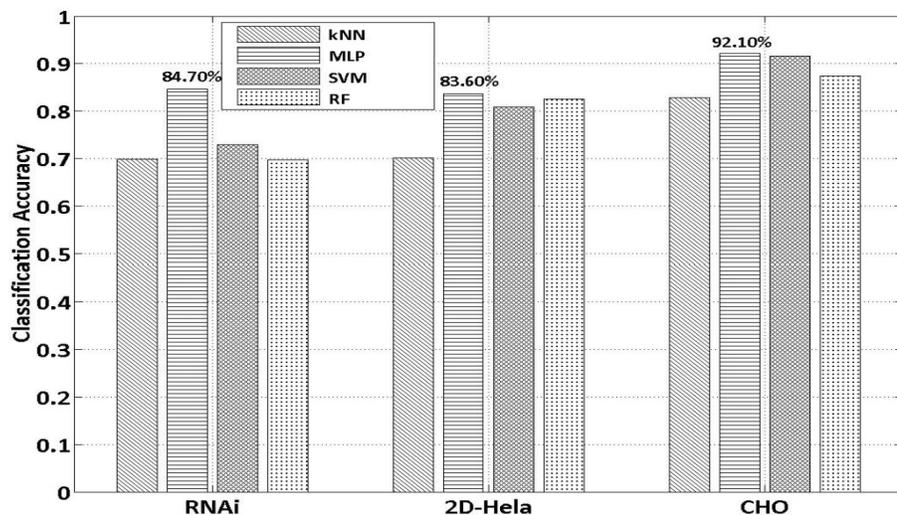


Figure 4: Barplots comparing the classification accuracies from the four classifiers.

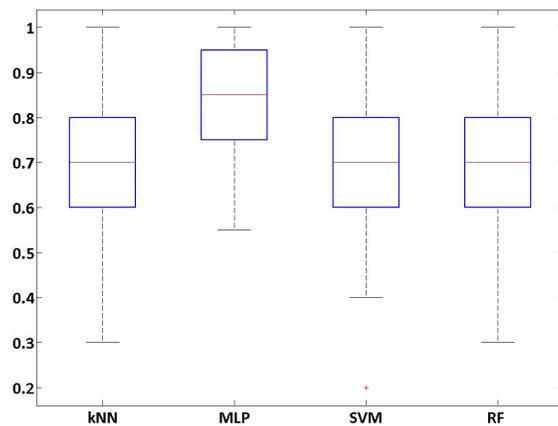


Figure 5: Boxplots comparing the classification accuracies from the four classifiers.

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.

Figure 4 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 84.7%, which is better than the published result 82% [12]. The accuracies from other three classifiers are 69.9% (k NN), 69.7% (random forest), and 72.9% (SVM). For 2D HeLa and CHO, the best results obtained by MLP are 83.6% and 92.1%, 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 [12]. 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 experiment is that MLP as a memory-based classifier is more resistant to insufficient data amount comparing the margin or distance-based SVM.

Figure 5 presents the box plot of classification results obtained by these four single classifiers on RNAi data set, it can be seen that the MLP classifier has the smallest variance range in classification result, the lowest classification rate of MLP is about 55%, which is much higher than the lowest classification rates of other three classifiers.

In the next experimental part of this study, we seek to prove 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 experiment was also carried out on the three fluorescence microscopy image data sets introduced in Section 2.

The settings for the experiment 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 testing and 10% for validation, 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

Table 1: Improvement of classification accuracy by using Random Subspace Ensemble

| Classifier | RNAi | 2D Hela | CHO |
|----------------------------|--------|---------|--------|
| MLP | 84.70% | 83.20% | 91.20% |
| MLP-RSE (ensemble size=20) | 85.6% | 85.91% | 93.2% |

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 20 components can be seen in Table 1.

From Table 1, one can see that for all the three 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, from 84.7% upgraded to 85.6%. The classification accuracy for the other two data sets also be improved as well, for 2D-hela, it has been enhanced from 83.6% to 85.91%; for CHO, the classification accuracy has been upgraded to 93.2%.

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 ensembles from 10 components networks to 80 networks in each of the ensemble. The experiment procedure is same as above with results of the averaged classification accuracies shown in Figure 6. It seems that bigger ensemble size does bring better classification performance for all of three image data sets. But such improvement becomes marginal after the size exceed a limit and the bigger ensemble sizes bring heavy computational burden on the training phase. As can be seen from Fig. 6, at the ensemble size 70, for all of the three image data sets, we reach better classification accuracies than other ensemble sizes. The classification results of these three data sets are enhanced comparing to the results in Shamir et al. [12].

In Table 2, Table 3 and Table 4, we only listed the top-10 classification results from single MLPs for comparsion, an apparent 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 linear perceptron, and the ensemble offers relatively smaller standard deviations.

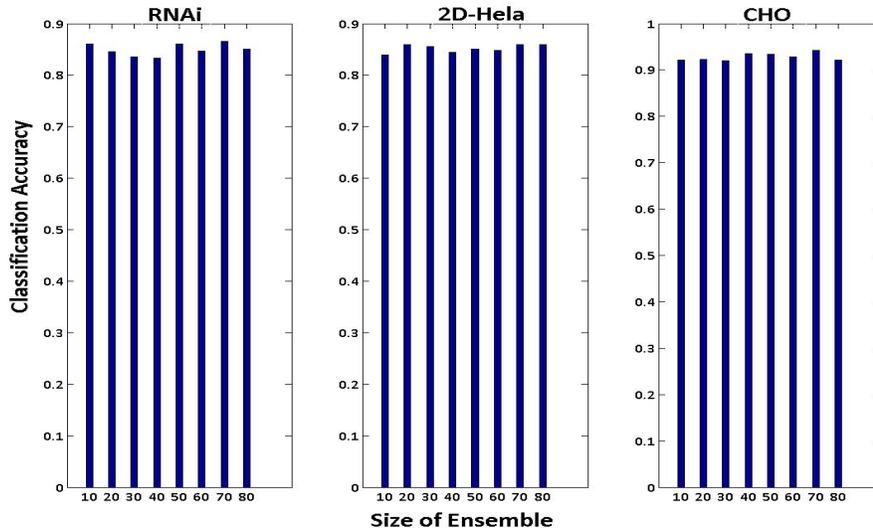


Figure 6: Classification accuracies on different ensemble sizes

Table 2: Performance from Random Subspace Ensemble of RNAi

| Indices | Accuracy (mean) | Standard Deviation |
|----------|-----------------|--------------------|
| 1 | 86.20% | 0.1270 |
| 2 | 86.00% | 0.1247 |
| 3 | 85.90% | 0.1280 |
| 4 | 85.90% | 0.1215 |
| 5 | 85.90% | 0.1248 |
| 6 | 85.90% | 0.1319 |
| 7 | 85.80% | 0.1232 |
| 8 | 85.80% | 0.1241 |
| 9 | 85.70% | 0.1148 |
| 10 | 85.60% | 0.1274 |
| Ensemble | 86.50% | 0.1282 |

In the following, we evaluated three different types of MLP ensembles for the RNAi, 2D-Hela and CHO image classification. The ensemble methods we compared are Random Subspace, Rotation Forest and Dynamic Classifier Selection. Rotation forest ensemble and dynamic classifier selection are two ensemble methods proposed recently, details of these two methods can be seen in [31] and [41]. The experiment settings for these three ensemble methods are

Table 3: Performance from Random Subspace Ensemble of CHO

| Indices | Accuracy (mean) | Standard Deviation |
|----------|-----------------|--------------------|
| 1 | 93.65% | 0.0607 |
| 2 | 93.53% | 0.0588 |
| 3 | 93.35% | 0.0577 |
| 4 | 93.35% | 0.0583 |
| 5 | 93.35% | 0.0595 |
| 6 | 93.35% | 0.0595 |
| 7 | 93.29% | 0.0620 |
| 8 | 93.18% | 0.0559 |
| 9 | 93.12% | 0.0580 |
| 10 | 93.12% | 0.0643 |
| Ensemble | 94.24% | 0.0542 |

Table 4: Performance from Random Subspace Ensemble of 2D-Hela

| Indices | Accuracy (mean) | Standard Deviation |
|----------|-----------------|--------------------|
| 1 | 85.12% | 0.0503 |
| 2 | 84.86% | 0.0492 |
| 3 | 84.79% | 0.0570 |
| 4 | 84.72% | 0.0547 |
| 5 | 84.70% | 0.0486 |
| 6 | 84.63% | 0.0541 |
| 7 | 84.63% | 0.0547 |
| 8 | 84.60% | 0.0527 |
| 9 | 84.58% | 0.0489 |
| 10 | 84.58% | 0.0543 |
| Ensemble | 85.96% | 0.0498 |

similar, the experiment procedure is same as above with results of the averaged classification accuracies. Figure 7 shows the comparison result of these three ensemble methods on ensemble size is 70, since in this size, we obtained the best classification accuracy.

Although in Figure 7, we only listed the results of ensemble size 70, in our experiment we found that for all the ensemble sizes we tested, random subspace ensemble always surpasses rotation forest. In the cases that the ensemble

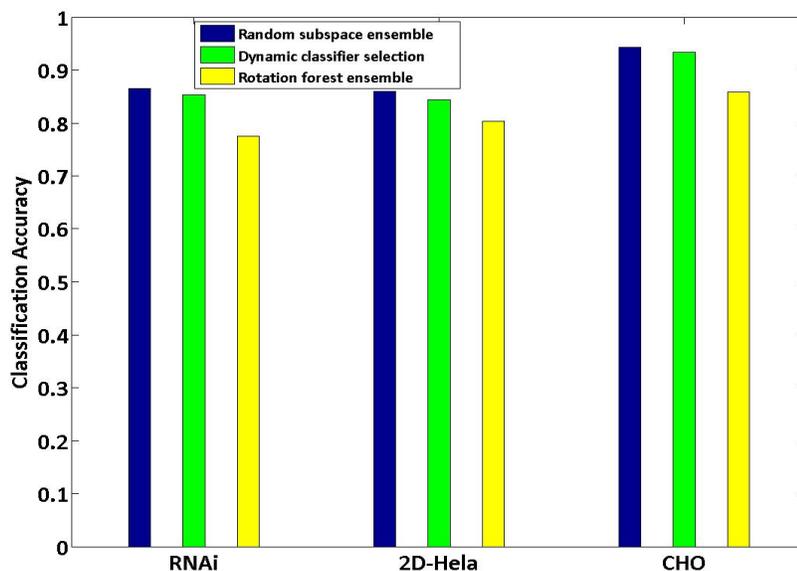


Figure 7: Comparison of different ensemble methods with ensemble size 70

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 [36].

The confusion matrices that summarize the details of the above random subspace ensemble on RNAi image data set is given in Table 5. For the total number of 10 testing samples (one for each category) in each experiment, the 10-by-10 matrix records the number of correct and incorrect predictions made by the classifier ensemble compared with the actual classifications in the test data. The matrices 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 and CHO data sets are given in Table 6 and Table 7, respectively.

Table 5: 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) | 0.07 | 0 | 0 | 0.8 | 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 | 0.13 | 0.19 | 0 | 0.02 | 0 | 0.61 |

Table 6: 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 | 0.74 | 0.03 | 0 | 0 | 0.19 | 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 | 0.16 | 0.78 | 0.04 | 0 | 0 | 0.02 | 0 |
| 6 (Golgpp) | 0 | 0.06 | 0 | 0 | 0.04 | 0.86 | 0 | 0.03 | 0.01 | 0 |
| 7 (Lysosome) | 0.02 | 0.05 | 0.05 | 0 | 0 | 0.03 | 0.8 | 0.04 | 0 | 0.01 |
| 8 (Microtubules) | 0 | 0.08 | 0.07 | 0.04 | 0 | 0.03 | 0.02 | 0.76 | 0 | 0 |
| 9 (Mitochondria) | 0 | 0 | 0 | 0.02 | 0 | 0 | 0 | 0 | 0.98 | 0 |
| 10 (Nucleolus) | 0 | 0 | 0.03 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0.95 |

Table 7: Averaged confusion matrix for CHO.

| % | 1 | 2 | 3 | 4 | 5 |
|-------------|------|------|------|------|------|
| 1 (Giantin) | 0.91 | 0 | 0.09 | 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 | 0.08 | 0.86 |

6 Conclusion

Classifiers ensemble 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 paper, we aimed at solving the challenging multi-class phenotype classification problem from microscopy images for RNAi screening. Two contributions are presented. Firstly, we proposed to apply the curvelet transform to efficiently describe microscopy 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 microscopy 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 benchmarking RNAi datasets showed that the random subspace MLP ensemble method achieved significantly higher classification accuracies ($\sim 86.5\%$), compared to the published result 82%. And the classification results of other two groups of microscopy image data sets using random subspace MLP also support the effectiveness of the proposed method.

ACKNOWLEDGEMENTS.

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

References

- [1] J. Wang, X. Zhou, PL. Bradley, SF.Chang, N.Perrimon, STC.Wong, Cellular Phenotype Recognition for High-Content RNA Interference Genome-Wide Screening. *J Biomol Screen*, **13**, (2008), 29-39.
- [2] M.Boutros, AA. Kiger, A. Armknecht, K. Kerr, M. Hild, R. Koch, SA. Haas, R. Paro, and N. Perrimon, Genome-wide RNAi analysis of growth and viability in drosophila cells. *Science*, **303**, (2004), 832-835.
- [3] ZE.Perlman, MD. Slack, Y.Feng, TJ. Mitchison, LF.Wu, SJ.Altshule, Multidimensional drug profiling by automated microscopy. *Science*, **306**, (2004), 1194-1198.

- [4] JC. Yarrow, Y.Feng, ZE.Perlman, T.Kirchhausen, T.J.Mitchison, Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb Chem High Throughput Screen*, **6**, (2003), 279-286.
- [5] MT.Weirauch, CK.Wong, AB. Byrne, and JM. Stuart, Information-based methods for predicting gene function from systematic gene knock-downs, *Bioinformatics*, **9**, (2008),463 doi:10.1186/1471-2105-9-463.
- [6] JC.Clemens, CA. Worby, N. Simonson-Leff, M. Muda, T. Maehama, BA. Hemmings, and JE. Dixon, Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. *Proc. Natl. Acad. Sci.*, **97**, (2000), 6499-6503.
- [7] GJ. Hannon, RNA interference. *Nature*, **418**, (2002), 244-251.
- [8] A.Kiger, B. Baum, S. Jones, M. Jones, A. Coulson, C. Echeverri, and N.Perrimon, A functional genomic analysis of cell morphology using RNA interference, *J. Biol.*, **2**, (2003), 27.
- [9] CJ. Echeverri and N. Perrimon, High-throughput RNAi screening in cultured cells: a user's guide. *Nat. Rev. Genet.*, **7**, (2006), 373-384.
- [10] N. Orlov, J. Johnston, T. Macura, L. Shamir, and I.Goldberg, Computer Vision for Microscopy Applications. *Vision Systems: Segmentation and Pattern Recognition*, Edited by: Goro Obinata and Ashish Dutta, pp.546, I-Tech, Vienna, Austria, June 2007.
- [11] H. Peng, Bioimage informatics: a new area of engineering biology. *Bioinformatics*, **24**(17), (2008), 1827-36.
- [12] L. Shamir, N. Orlov, DM. Eckley, T.Macura, I. Goldberg, IICBU 2008 - a proposed benchmark suite for biological image analysis, *Medical & Biological Engineering & Computing*, **46**, (2008), 943-947.
- [13] MV.Boland and RF. Murphy, A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, **17**, (2001), 1213-1223.

- [14] MV.Boland, M.Markey and RF. Murphy, Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images. *Cytometry*, **33**, (1998), 366-375.
- [15] K. Huang and RF. Murphy, Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics*, **5**, (2004), 78.
- [16] A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, RF. Murphy and J.Kovacevic, A multiresolution approach to automated classification of protein subcellular location images. *Bioinformatics*, **8**, (2007), 210.
- [17] L.Nanni, A. Lumini, Y. Lin, C. Hsu, and C. Lin, Fusion of systems for automated cell phenotype image classification. *Expert Systems with Applications*, **37**, (2010), 1556-1562.
- [18] NA. Hamilton, RS. Pantelic, K. Hanson and RD. Teasdale, Fast automated cell phenotype image classification. *Bioinformatics*, **8**, (2007), 110.
- [19] J.Starck, E. Candès and DL. Donoho, The Curvelet transform for image denoising, *IEEE Transactions Image Processing*, **11**, (2002), 670-684.
- [20] E. Candès and DL. Donoho, Curvelets - a surprisingly effective nonadaptive representation for objects with edges, in *Curves and Surfaces*, Vanderbilt University Press, 1999.
- [21] E. Candès, L. Demanet, DL. Donoho and L. Ying, Fast discrete curvelet transforms, *Multiscale Modeling and Simulation.*, **5**, (2006), 861-899.
- [22] T. Gebäck and P. Koumoutsakos, Edge detection in microscopy images using curvelets, *BMC Bioinformatics*, **10**, (2009), doi:10.1186/1471-2105-10-75.
- [23] L. Ni and H. Leng, Curvelet transform and its application in image retrieval, in *3rd International Symposium on Multispectral Image Processing and Pattern Recognition, Proceedings of SPIE*, **5286**, (2003), 927-930.
- [24] A. Majumdar and R. Ward, Multiresolution methods in face recognition, In *Recent Advances in Face Recognition*, Kresimir Delac, Mislav Grgic and

- Marian Stewart Bartlett (Ed.), ISBN: 978-953-7619-34-3, InTech, Croatia, 2008.
- [25] T. Mandal, QJ. Wu and Y. Yuan, Curvelet based face recognition via dimension reduction, *Signal Processing*, **89**, (2009), 2345–2353.
- [26] I. Sumana, M. Islam, D. Zhang and G. Lu, Content based image retrieval using curvelet transform, in *2008 IEEE 10th Workshop on Multimedia Signal Processing*, (2008), 11-16, Cairns, Australia.
- [27] L.Hansen, P. Salamon, Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, (1990), 993-1001.
- [28] LI.Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004.
- [29] L. Breiman, Bagging predictors, *Machine Learning*, **24**, (1996), 123-140.
- [30] Y.Freund, RE. Schapire, A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, **55**, (1997), 119-139.
- [31] J.Rodríguez, LI. Kuncheva and C. Alonso, Rotation Forest: A New Classifier Ensemble Method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, (2006), 1619-1630.
- [32] S. Haykin, *An Introduction to Neural Networks: A Comprehensive Foundation, 2nd ed.* Upper Saddle River, NJ: Prentice-Hall, 2000.
- [33] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis.* Cambridge University Press, 2004.
- [34] L. Breiman, Random Forests. *Machine Learning*, **45**(1), (2001), 5-32.
- [35] TK. Ho, The Random Subspace Method for Constructing Decision Forest, *IEEE Trans PAMI*, **20**, (1998), 832-844.
- [36] M. Skurichina and R. P. W. Duin, Bagging, Boosting and the Random Subspace Method for Linear Classifiers, *Pattern Analysis & Applications*, **5**, (2002), 121-135.

- [37] L. I. Kuncheva, J. J. Rodríguez and C. O. Plumpton, Random Subspace Ensembles for fMRI Classification, *IEEE Transactions on Medical Imaging*, **29**, (2010),531-542.
- [38] A. Bertoni, R. Folgieri and G. Valentini, Bio-molecular cancer prediction with random subspace ensembles of Support Vector Machines, *Neurocomputing*, **63**, (2005), 535-539.
- [39] A. H. Ko, R. Sabourin, and A. de Souza Britto, Combining Diversity and Classification Accuracy for Ensemble Selection in Random Subspaces, in *2006 International Joint Conference on Neural Networks*, (2006), 2144-2151 , Vancouver, BC , Canada.
- [40] A. H. Ko, R. Sabourin, and A. de Souza Britto, Evolving Ensemble of Classifiers in Random Subspace, in *Proceedings of the 8th annual conference on Genetic and evolutionary computation, 2006*, (2006), 1473-1480, Seattle, Washington, USA.
- [41] K. Woods, W. P. Kegelmeyer and K. Bowyer, Combination of Multiple Classifiers Using Local Accuracy Estimates, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, (1997), 405-410.