Journal of Theoretical Biology 340 (2014) 85–95

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Protein subcellular localization in human and hamster cell lines: Employing local ternary patterns of fluorescence microscopy images

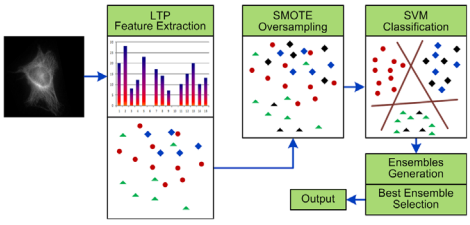
Muhammad Tahir [a], Asifullah Khan [a,*], Hüseyin Kaya [b]

[a] Department of Computer and Information Sciences, PIEAS, Islamabad, Pakistan
[b] Department of Biophysics, Faculty of Medicine, University of Gaziantep, 27310 Gaziantep, Turkey

HIGHLIGHTS

- LTPs exploit small variations in intensities of Human and Hamster protein images.
- SMOTE oversampling is utilized to increase the minority class samples.
- SVM shows significance performance improvement for balanced data.
- mRMR is not required for the performance improvement of LTPs.
- A web server is available online at http://111.68.99.218/Protein_SubLoc.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 May 2013
Received in revised form
9 July 2013
Accepted 15 August 2013
Available online 27 August 2013

Keywords:
Support vector machine
SMOTE
mRMR

ABSTRACT

Discriminative feature extraction technique is always required for the development of accurate and efficient prediction systems for protein subcellular localization so that effective drugs can be developed. In this work, we showed that Local Ternary Patterns (LTPs) effectively exploit small variations in pixel intensities; present in fluorescence microscopy based protein images of human and hamster cell lines. Further, Synthetic Minority Oversampling Technique is applied to balance the feature space for the classification stage. We observed that LTPs coupled with data balancing technique could enable a classifier, in this case support vector machine, to yield good performance. The proposed ensemble based prediction system, using 10-fold cross-validation, has yielded better performance compared to existing techniques in predicting various subcellular compartments for both 2D HeLa and CHO datasets. The proposed predictor is available online at: http://111.68.99.218/Protein_SubLoc/, which is freely accessible to the public.

1. Introduction

Protein is the crucial part of a cell in all living organisms to function properly. Among numerous characteristics, subcellular localization is the most important property of proteins (Chebira et al., 2007). Understanding the behaviour of individual protein is the key to the comprehension of various functions of cells in living organisms. A protein must reside in its natural localization to work properly. Hence, precise knowledge of protein subcellular localization allows to elucidate various protein functions (Lin et al., 2007). In addition, various cellular processes of hypothetical and newly revealed proteins can easily be described by the protein localization (Boland and Murphy, 2001; Murphy et al., 2000). Further, subcellular localization can aid in drug discovery (Nanni and Lumini, 2008). For instance, plasma membrane proteins and secreted proteins are easily reachable by drug molecules since they are located on the cell surface (Tscherepanow et al., 2008). Subcellular localization also helps in early diagnostics of various

* Corresponding author at: Postal address: Pattern Recognition Laboratory, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore 45650, Islamabad, Pakistan.
Tel.: +92 51 2207380-84; fax: +92 51 2208070.
E-mail addresses: asif@pieas.edu.pk, khan.asifullah@gmail.com (A. Khan).

0022-5193/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.jtbi.2013.08.017

diseases. For example, aberrant subcellular localization has been observed in the cells affected by various diseases such as Alzheimer′s and cancer. In addition, the environment in which proteins function properly can also be determined by accurately finding their subcellular localization (Chen et al., 2006). In brief, accurate knowledge of protein localization is helpful in the identification as well as effectiveness of drugs.

Experimental classification of protein localization is time consuming and laborious (Murphy et al., 2000). Therefore, computational methods coupled with machine learning techniques are required to determine protein subcellular localizations. In this regard, researchers have endeavoured to develop numerous bioinformatics based prediction systems coupled with machine learning methods to localize a range of proteins (Chebira et al., 2007; Hamilton et al., 2007; Khan et al., 2008, 2011; Lin et al., 2007; Murphy et al., 2003; Nanni and Lumini, 2008; Nanni et al., 2010a; Zhang et al., 2009). Researchers have confirmed that many proteins have been found to be the part of a multi-label system in which they are able to reside in two or more subcellular locations simultaneously or travel across two or more subcellular location sites. This property of proteins, making them unique in their biological functionality, is of particular interest (Chou, 2013). In this connection, substantial efforts have been endowed for the last three years to localize multiplex proteins in addition to the singleplex proteins. In this regard, various prediction systems have been developed focussing on different organisms including animal (Lin et al., 2013), human (Chou et al., 2012), bacterial proteins (Wu et al., 2012; Xiao et al., 2011a), plant (Wu et al., 2011), virus (Xiao et al., 2011b), and Eukaryotic Proteins (Chou et al., 2011).

From the literature survey, it is observed that sequence based methods cover more subcellular location sites compared to image-based methods. In addition, singleplex and multiplex proteins are mostly covered by sequence based techniques. For example, a benchmark dataset utilized in Chou et al. (2011) covers 22 subcellular locations. Similarly, another benchmark dataset reported in Lin et al. (2013) has 20 subcellular location sites. Likewise, 14 subcellular locations of human proteins are reported in Chou et al. (2012). Due to the wider coverage of protein location sites, the applications of sequence-based methods are more likely. However, the current work is based on the fluorescence microscopy images; therefore, the singleplex proteins are targeted in order to simplify the treatment. The proposed method covers 10 and 8 subcellular locations in HeLa and CHO datasets.

Literature survey has revealed that both individual and ensemble classifiers have been employed in conjunction with various feature extraction strategies to accurately predict subcellular localization (Chen et al., 2006; Hu and Murphy, 2004). A model has been developed in which a random subspace of local binary and ternary patterns with high variance is selected. Reduced dimensionality is achieved through Neighbourhood Preserving Embedding. Further, support vector machine (SVM) is trained using the reduced dimensionality space (Nanni et al., 2010b). Similarly, an adaptive multi-resolution approach has been proposed in which Haralick textures and morphological features are extracted at the sub-bands. The predictions at different sub-bands are obtained by utilizing *k*-means algorithm and weighting (Srinivasa et al., 2006). In another approach, a random subspace of Levenberg–Marquardt neural networks and a variant of the AdaBoost learning algorithms are trained using hybrid feature sets. The decisions of the two ensembles are fused together through sum rule (Nanni et al., 2010c). Similarly, a model based on back propagation neural network has been reported, which employs Haralick textures, Zernike moments, and morphological features for protein subcellular localization (Murphy et al., 2003). Likewise, an SVM based model is proposed, which utilizes Zernike moments, Threshold Adjacency Statistics (TASs), and hybrid feature space of TASs and Haralick textures (Hamilton et al., 2007). An Artificial Neural Network based prediction system has been proposed, which utilizes Haralick, morphological and Zernike moment based features in multi-resolution subspaces. Final decision is made through weight assignment (Chebira et al., 2007).

The existing approaches have shown great improvement in achieving higher accuracies; however, we need systems, which are capable of achieving nearly 100% accuracy particularly, in diagnosing cancer like diseases. In addition, the nature of available unbalanced data should also be taken into account to enhance the performance of the classifier. In our previous work, we have focussed on the same problem exploiting different spatial and transform domain features (Tahir et al., 2012). The ensemble classification based on different SVM kernels has achieved 99.7% accuracy for the 2D HeLa dataset. However, we did not take into account the unbalanced nature of the data. The principal focus of this study is to develop a model that is reliable, efficient, and highly accurate even in the presence of unbalanced data keeping the feature space as small as possible. In order to have small and well discriminative feature space, Local Ternary Patterns (LTPs) (Nanni et al., 2010b) have been utilized for the classification of protein subcellular localization images. Further, Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) has been employed to balance the datasets. The feature selection technique; minimum Redundancy Maximum Relevance (mRMR) has also been utilized to know whether the feature space has any overlapping information. Polynomial SVM of degree 2 has been employed to nonlinearly transform the input features to make them linearly separable. The proposed novel combination of LTPs, SMOTE, and SVM performed well for protein classification compared to the existing ensemble classification techniques, especially, due to the exploitation of discriminative capability of LTPs as well as availability of balanced feature space constructed using SMOTE. The discrimination power of LTPs is better because of its less sensitive behaviour in homogenous image regions.

The rest of the paper is structured as follows. Section 2 discusses materials and methods. Section 3 describes the performance measures used in this paper. Section 4 is devoted to the analysis of experimental results. Section 5 draws the conclusion at the end.

2. Materials and methods

In this section, we present the datasets, the feature extraction and post-processing techniques as well as the classifier adopted to develop the proposed model.

2.1. Datasets

We have tested the performance of proposed model using 2D HeLa and Chinese Hamster Ovary (CHO) datasets from Murphy′s Lab (Murphy, 2004) and AIIA lab (Lin et al., 2007), respectively. In 2D HeLa dataset, there are 862 images distributed in ten distinct classes including ActinFilaments, Endosome, ER, Golgi Giantin, Golgi GPP130, Lysosome, Microtubules, Mitochondria, Nucleolus, and Nucleus. Sample image from each class is shown in Fig. 1. Image distribution, per class in the original and oversampled datasets, is illustrated in Fig. 2.

On the other hand, CHO dataset has 668 protein images grouped in eight different categories, which include Actin, ER, Golgi, Microtubule, Mitochondria, Nucleolus, Nucleus, and Peroxisome. Some sample images are illustrated in Fig. 3. Comparison of the synthetic samples and original samples are shown in Fig. 4.

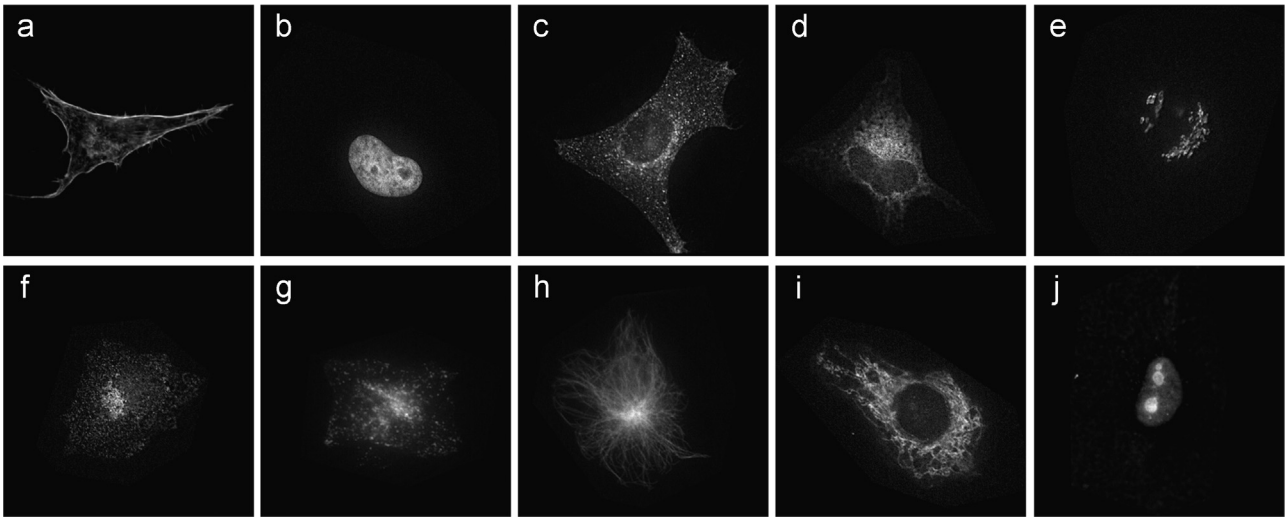


Fig. 1. (a)–(j) HeLa images, one for each class, (a) ActinFilaments (b) Nucleus (DNA) (c) Endosome (d) ER (e) Golgi Giantin, (f) Golgi GPP130 (g) Lysosome (h) Microtubules (i) Mitochondria (j) Nucleolus.

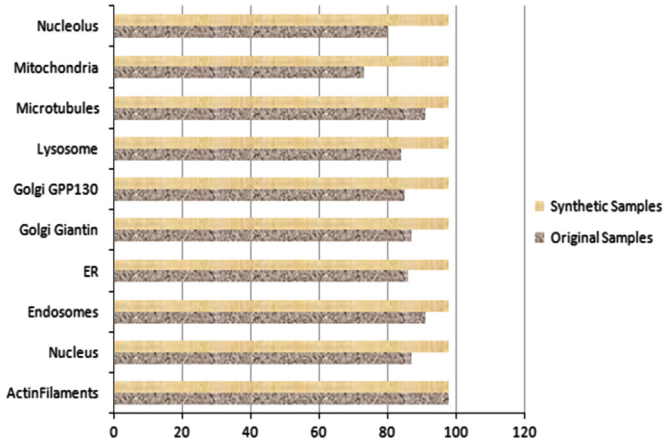


Fig. 2. Comparison of original and synthetic samples in 2D HeLa dataset.

2.2. The proposed system

The framework of our proposed system is demonstrated in Fig. 5. In the first stage of the proposed prediction system, LTPs of each image are extracted. Example instances are shown in different colours. Next, SMOTE has been applied in order to synthetically increase the samples of each minority class. The oversampled instances are shown in black. The resultant oversampled features are provided as input to the classification stage, where SVM is trained as classifier. We have utilized six different LTPs features, hence, obtained six SVM classifiers. Among these classifiers, we have chosen five classifiers, which produced the highest performance prediction for the ensemble generation. The ensemble output is made through the majority voting technique.

mRMR has also been utilized to select the most appropriate features for the classification stage. SVM is trained as classifier, however, mRMR based features have not shown improvement over the original feature space.

2.3. Feature extraction

In the feature extraction stage, we represent an image using attribute values in the form of a feature vector. For this purpose, we usually adopt a feature extraction technique. In this section, we discuss LTPs, which is utilized as a feature extraction technique in this work. LTPs (Nanni et al., 2010b) is a texture based feature extraction technique, which is less sensitive to noise compared to other texture based feature extraction techniques. On the basis of negative and positive components, each LTPs is split into two Local Binary Patterns (LBPs) (Nanni et al., 2010b) in order to reduce the computational complexity of LTPs as depicted in Fig. 6. The two histograms, obtained from the computation of constituent LBPs, are concatenated together to obtain the final feature vector of LTPs. In calculating LTPs, the relation of a central pixel c with its neighbour u is defined by the difference operation.

The encoding of the LTPs is performed when one of the three conditions is met according to a given threshold θ as:

$$s(u) = \begin{cases} 1 & \text{if } u \geq c+\theta \\ -1 & \text{if } u \leq c-\theta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

LTPs are usually computed using three distinct mappings including uniform mapping, rotation invariant mapping, and uniform rotation invariant mapping. LTPs are computed over three different configurations i.e. ($R=1$, $N=8$), ($R=2$, $N=16$), and ($R=3$, $N=24$), where R and N represent radius and number of neighbouring pixels, respectively.

2.4. Post-processing

In the post processing step, we usually perform certain operations in order to refine the feature space for the classification step. In this work, we have employed two post-processing techniques; SMOTE and mRMR. The former is mandatory in our proposed approach while the later is optional.

The performance of the classifier may be degraded due to the unbalanced nature of training data. The majority class gains the favor of the classifier whereas the performance of the minority class is greatly affected. Therefore, we have utilized an oversampling technique in order to increase the samples of minority classes so that they have the same number of samples as there are in the majority class. We have adopted SMOTE (Chawla et al., 2002) to reduce the bias produced due to the unbalanced nature of data. SMOTE performs its operations in the feature space. SMOTE produces synthetic samples of minority classes so that the classifier produces a decision surface that is more generalized. In 2D HeLa dataset, class 1 has the largest number of samples in the dataset as shown in Fig. 2 whereas the remaining classes have less

number of samples. In this work, SMOTE has been employed to add synthetic samples to all the nine minority classes.

In order to introduce synthetic samples for the minority class in the feature space, SMOTE selects a minority class sample and creates novel synthetic samples along the line segment joining some or all k nearest neighbors belonging to that class. SMOTE does not replicate the original sample for oversampling; instead it follows a unique procedure to generate new samples. SMOTE subtracts the selected sample of the minority class from its nearest neighbor and then multiplies the output of this step by a random number ranging from 0 to 1. The new sample is obtained by adding this result to the originally selected sample of the minority class. This synthetic instance is assured to be on the line segment separating two specific features.

In this way, the classifier learns more general decision regions for the samples belonging to minority class due to the newly introduced synthetic samples. Thus minority class gets more attention of the classifier and bias towards the majority class is reduced.

Feature selection is a significant requirement for designing classification models in the field of pattern recognition and classification. In feature selection, an optimal subset of the whole feature space is utilized having the same discriminative information as the original set. The core objective of feature selection is to reduce the dimensionality of the feature space while selecting those features that have at least the same discriminative power as the original set. This procedure should lead to the non-redundant feature set that has reduced noise and less computational cost. In this work, we have adopted mRMR as a feature selection technique, which has been employed by various researchers in the field of bioinformatics and machine learning (He et al., 2010, 2012; Huang et al., 2010, 2011, 2013; Li et al., 2012a, 2012b, 2012c; Yi et al., 2012). mRMR takes into account the relevance and redundancy scores of each feature. The mRMR selected feature space bears minimum redundancy with other features and maximum relevance to the target class. The mutual information among the features themselves as well as the features and the class variables can be utilized to calculate the correlation and relevance. The mutual information among the features can be estimated using Eq. (2).

$$MI(x,y) = \sum_{i,j \in N} p(x_i,y_j)(\log\ p(x_i,y_j)/p(x_i)p(y_j)) \tag{2}$$

here x and y represent two features, $p(x_i,y_j)$ shows joint probabilistic density function and $p(x_i)p(y_j)$ indicates marginal probabilistic density function. Likewise, the mutual information between the features and the target class variables are computed using Eq. (3).

$$MI(x,z) = \sum_{i,k \in N} p(x_i,z_k)(\log\ p(x_i,z_k)/p(x_i)p(z_k)) \tag{3}$$

here x represents a feature and z indicates a target class.

The minimum redundancy in the entire feature space is computed using Eq. (4).

$$\min(mR) = (1/|S|^2) \sum_{x,y \in S} MI(x,y) \tag{4}$$

in this equation, S stands for the feature space and $|S|$ shows total number of features in the feature space.

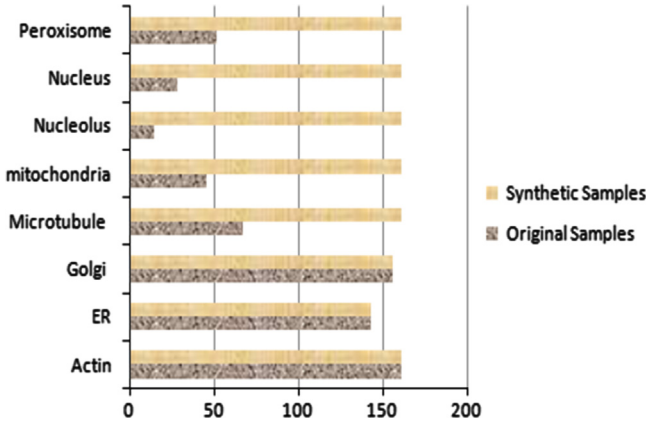


Fig. 4. Comparison of original and synthetic samples in CHO dataset.

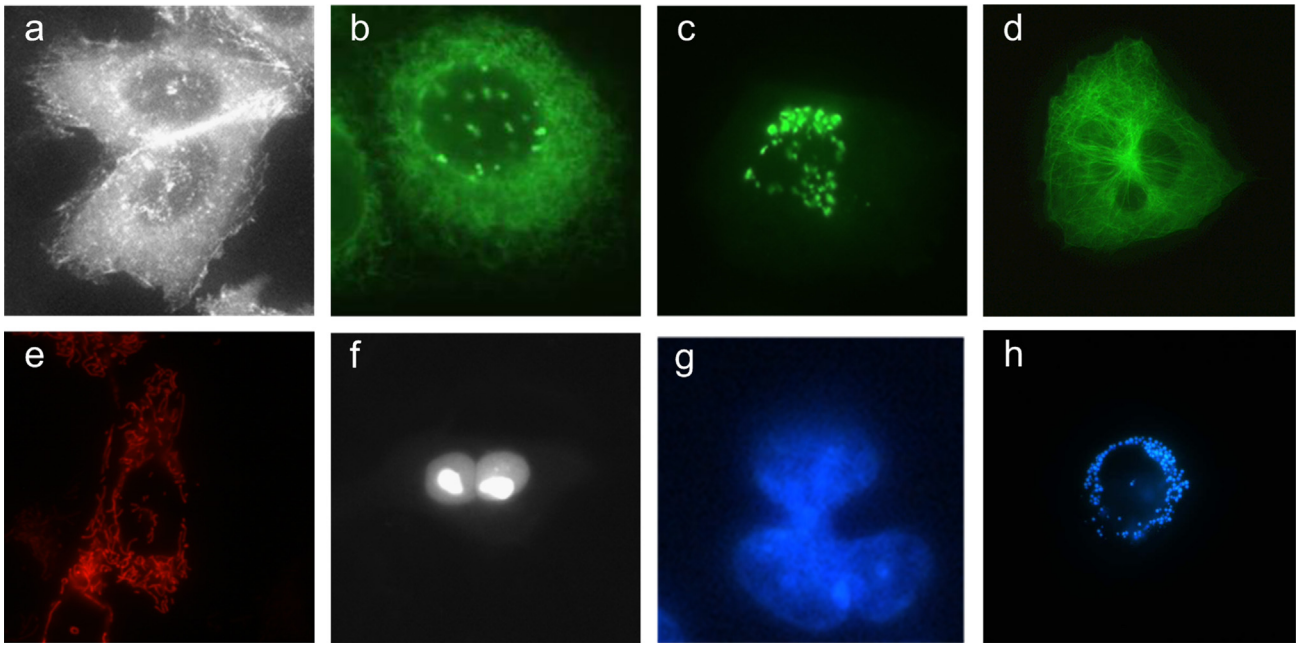


Fig. 3. (a)–(j) CHO images, one for each class, (a) Actin (b) ER (c) Golgi (d) Microtubule, (e) Mitochondria (f) Nucleolus (g) Nucleus (h) Peroxisome.

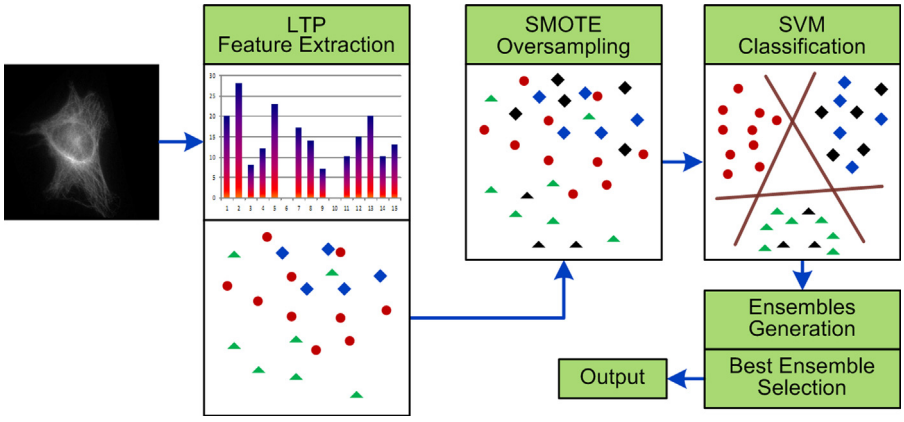


Fig. 5. Framework of the proposed system.

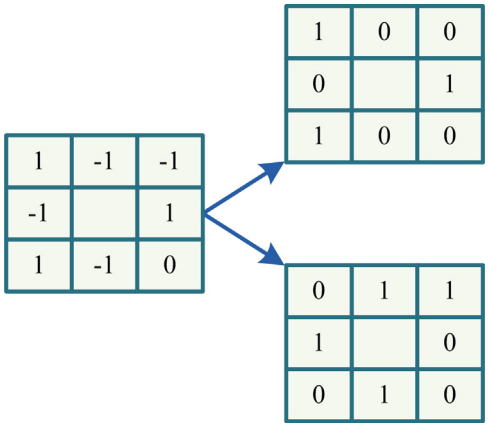


Fig. 6. Formation of Two LBP codes from single LTPs code.

The maximum relevance between the features and target class variables is obtained using Eq. (5)

$$\max(MR) = (1/|S|) \sum_{x \in S} MI(x, z) \tag{5}$$

Through the simultaneous optimization of Eqs. (4) and (5), the final feature space is obtained using equation:

$$\max(\nabla MI) = MR - mR$$

2.5. Polynomial SVM based classification

The computed feature vectors, in the feature extraction step, are delivered to the classifier so that it can group the instances in different classes. In this work, we have utilized SVM with polynomial kernel of degree 2 as classification algorithm to nonlinearly transform the features so they could be made linearly separable.

Theoretical details of SVM can be found in the statistical and machine learning theory (Gunn, 1998; Hayat et al., 2011). It has been used in the field of bioinformatics, computational biology, and pattern recognition (Hayat and Khan, 2011, 2012; Khan et al., 2011; Li et al., 2003; Shi et al., 2007; Tahir et al., 2012; Zhang and Ding, 2007). SVM classifier builds a decision surface that maximizes the separation to the nearest data samples in the training data. SVM comes across with an optimal linear hyperplane, which has minimum classification error on new test samples. In order to classify linearly separable data points, SVM constructs a hyperplane such that the distance between the support vectors is maximized.

Let we have N training pairs (x_i, y_i), where x and y denote the data sample and its label, respectively. The functional form of a decision surface for linearly separable data is given by:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i x_i^T . x + bias \tag{6}$$

here, $\alpha_i > 0$ is the Langrange multiplier. The pattern vector x_i corresponding to $\alpha_i > 0$ is termed as support vector. Function $f(x)$ is not affected by the dimension of the feature space.

The function of a decision surface for non-linearly separable patterns is obtained using Eq. (7) as:

$$\Psi(w, \zeta) = (1/2) w^T w + C \sum_{i=1}^{N} \zeta_i \tag{7}$$

provided that the following condition is true: $y_i(w^T \Psi(x_i) + bias) \geq 1 - \zeta_i, \zeta_i \geq 0$

In Eq. (7), C is the penalty parameter of the error term, and $\Psi(x)$ is the nonlinear mapping. The weight vector w controls the cost function term $w^T w$ and is used here to minimize this term.

For non-linearly separable training data, lower dimensional data is mapped to a higher dimension using the mapping function $\Psi(x)$ provided that the condition is fulfilled as given: $\Psi : R^N \rightarrow F^M, M >> N$.

A kernel function is used to define each data point in the new feature space as given below:

$$K(x_i, x_j) = \Psi(x_i) \times \Psi(x_j) \tag{8}$$

In this work, we have used polynomial kernel of degree 2. It can be formulated as:

$$K(x_i, x_j) = ((\Psi(x_i) \times \Psi(x_j)) + 1)^d \tag{9}$$

here, d is the kernel parameter and indicates the degree of polynomial kernel, which is used to manipulate the decision boundary. The complexity of the decision surface increases in the input space (transformed feature space) by raising the degree of the polynomial kernel. Therefore, one can easily control the flexibility of the classifier by increasing the degree of the polynomial kernel. Polynomial kernel is useful in situations where the dimensionality of the feature space is low and number of instances is relatively higher. In this work, the feature space is normalized, dimensionality of the feature space is not high and number of instances is more compared to the feature space dimensions, therefore, polynomial SVM is the best choice among the available kernels. Further, from our previous studies (Tahir et al., 2012), we have found that polynomial SVM performed better compared to other kernels for these particular LTPs features.

3. Performance parameters

In machine learning, various performance measures are in practice to evaluate the performance of a model. These performance measures are computed from a confusion matrix. We have used accuracy, sensitivity, specificity, MCC, and *F*-measure as performance parameters in this work. A brief introduction to the adopted performance measures is as follows.

3.1. Accuracy

Accuracy estimates the overall efficacy of the algorithm. It is given by Eq. (10)

$$\text{Accuracy} = ((TP+TN)/(TP+FP+FN+TN)) \times 100 \quad (10)$$

TP, FN, TN, and *FP* are the number of true positive, false negative, true negative, and false positive fluorescence microscopy images, respectively.

3.2. Sensitivity/specificity

Sensitivity approximates the actual proportion of positive samples, which are correctly predicted by the classifier whereas specificity assesses the actual proportion of negative samples, which are correctly predicted by the classifier.

$$\text{Sensitivity} = (TP/(TP+FN)) \times 100 \quad (11)$$

$$\text{Specificity} = (TN/(TN+FP)) \times 100 \quad (12)$$

3.3. MCC

MCC is a discrete version of Pearson's correlation coefficient that returns a scalar value in the range of -1 and $+1$; -1 means the classifier always makes a mistake whereas $+1$ means the classifier never makes a mistake.

$$MCC(i) = ((TP \times TN - FP \times FN)/(\sqrt{[TP+FP][TP+FN][TN+FP][TN+FN]})) \quad (13)$$

in order to understand the importance of MCC for measuring the quality of a predictor, readers are referred to study (Chen et al., 2013; Xu et al., 2013).

3.4. F-Measure

F-Measure is used to assess the accuracy of the performed test. *F*-Measure takes into account both the precision *p* and the recall *r* of the test. Precision is the number of true predictions divided by the number of all returned predictions whereas *recall* is the number of true predictions divided by the number of originally true predictions. *F*-Measure varies between 0 and 1. Close the returned value to 1 better is the *F*-measure.

$$F\text{-}measure = 2 \times ((precision \times recall)/)(precision + recall) \quad (14)$$

$$precision = (TP/(TP+FP)) \quad (15)$$

$$recall = (TP/(TP+FN)) \quad (16)$$

3.5. Q-Statistics

The average value of *Q*-statistic is used to measure the diversity in ensemble classifiers. In fact, the *Q*-statistic returns the similarity between any two classifiers. Therefore, the low value of *Q*-statistic shows the high diversity among the classifiers. In order to obtain the diversity value, we usually subtract the similarity output from 1.

The *Q*-statistic of any two base classifiers is defined as:

$$Q_{i,j} = \begin{cases} ((ad-bc)/(ad+bc)) & \text{if } a,\ b,\ c,\ \text{d} < 1 \\ 1 & otherwise \end{cases} \quad (17)$$

here *a,* and *d* represent the number of correct and incorrect prediction of both classifiers. However, *b* is the correct prediction of classifier first and incorrect prediction of classifier second and *c* is the correct prediction of classifier second and incorrect of first. The value of *Q* varies from -1 to $+1$, the lower the value of *Q* more diversity the classifier has. The classifiers that predict the same instance correctly would yield positive values of *Q*, and those classifiers that incorrectly classify different objects would yield negative values of *Q*. For statistically independent classifiers the value of $Q_{i,j}$ is zero (Meynet and Thiran, 2010). For ensemble classifier the average value of *Q*-statistic among all pairs of *L* base classifiers is calculated as:

$$Q_{avg} = (2/(L(L-1))) \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k} \quad (18)$$

3.6. Multiclass ROC

Area under curve (AUC) is a valuable performance metric, in terms of Receiver Operating Characteristic (ROC), for measuring the similarity between two different categories. We have adopted the method presented in Hand and Till (2001) to compute the AUC for 10 and 8 class classification problems. In this case, AUC is computed for all pair wise combinations of these 10 and 8 classes. Total AUC is the mean of all AUCs.

4. Results and discussion

In the literature, various statistical testing methods are available to evaluate the performance of a classification system. In order to assess the performance of a prediction system, three cross-validation techniques are generally adopted in the field of pattern recognition, machine learning, and bioinformatics. These include independent dataset test, sub-sampling test, and jackknife test. Among these, jackknife test is considered the best that always produces unique prediction outcomes for a given benchmark dataset as demonstrated by Eqs. 28–30 in Chou (2011). Therefore, researchers have been increasingly utilizing jackknife test in order to analyze the effectiveness of various predictors (Chen and Li, 2013; Chou, 2001; Esmaeili et al., 2010; Georgiou et al., 2009; Khosravian et al., 2013; Mei, 2012; Mohabatkar, 2010; Mohabatkar et al., 2011; Sahu and Panda, 2010; Zhang et al., 2008). The computational time of the jackknife test is reduced through the 10-fold cross-validation technique. In 10-fold cross validation, classifier is trained on 9/10 folds and tested on remaining 1/10 fold. This process is repeated ten times so that the classifier has gone through each fold for training and testing.

4.1. Prediction performance for 2D HeLa dataset

Table 1 presents the classification results of polynomial SVM of degree 2 using LTPs for 2D HeLa dataset. The very first column shows various mapping methods denoted by *m* through which these features are computed. These include rotation invariant, uniform, and uniform rotation invariant mappings. The second and third columns represent radius and neighborhood of operation, respectively.

Fourth column of Table 1 shows threshold values represented by θ at which these features are extracted. *D* is used to represent dimension of the feature space. Accuracies obtained are presented

in column six followed by sensitivity, specificity, MCC, and *F*-score in columns 7–10, respectively.

It is evident from the observation of Table 1 that all the LTPs have yielded accuracies above 90%. However, uniform rotation invariant LTPs outperformed all other LTPs. The highest accuracy achieved is 95.4% on the configuration of radius 3 in the neighborhood of 24 pixels. In order to better understand the effectiveness of SMOTE data balancing technique, we have also provided the simulation results using unbalanced HeLa cell lines as shown in Supplementary Table 5. The accuracy achieved using the balanced dataset is 1.4% higher than that of using the unbalanced dataset. For all the LTPs, SMOTE has positive effect on the performance of the prediction system.

High values of sensitivity and specificity revealed that the proposed model has maintained the balance in true positive and true negative predictions. Best prediction quality is observed against uniform rotation invariant LTPs on radius 3 where MCC value is 0.79. Similarly, test′s accuracy is also good as given by *F*-measure value of 0.80. In general, improved performance is observed on larger image patches.

The argument is further strengthened by the explained variance phenomenon. Explained variance is the ratio of sum of first two *PCA* components to the sum of all *PCA* components. Higher ratio indicates that the first two components contain most of the information. In this particular case, as depicted in Fig. 7, the ratio of explained variance to the total variance is 77%, which indicates that the first two components hold most of the information. However, it is evident that the first five components amount to 92% of the total variability present in the data as shown in Fig. 7. This means on using the first 5 components would certainly improve the performance of the classifier. We also computed multiclass *AUC* for uniform rotation invariant LTPs on radius 3 and neighborhood 24. The *AUC* value 0.73 shows good separation

Table 1
Performance of poly-SVM using LTPs for balanced HeLa dataset.

m	*R*	*N*	*θ*	*D*	Polynomial kernel				
					Acc	**Sen**	**Spe**	**MCC**	***F*-Score**
Rotation invariant	1	8	40	72	90.7	92.6	90.5	0.65	0.66
Uniform	1	8	40	118	93.1	93.7	93.0	0.71	0.73
Uniform rotation invariant	1	8	40	20	91.6	92.5	91.5	0.67	0.68
uniform	2	16	80	486	**95.4**	96.0	95.3	0.79	0.80
Uniform rotation invariant	2	16	80	36	94.6	95.9	94.5	0.77	0.78
Uniform rotation invariant	3	24	80	**52**	**95.4**	96.1	95.3	0.79	0.80

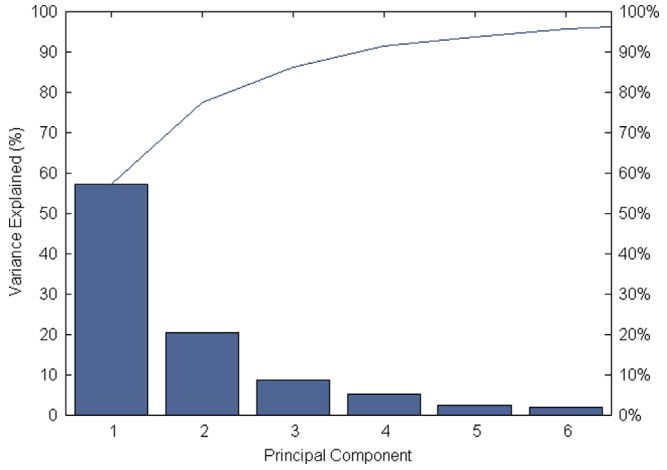


Fig. 7. Ratio of explained variance to the Total Variance for 2D HeLa dataset.

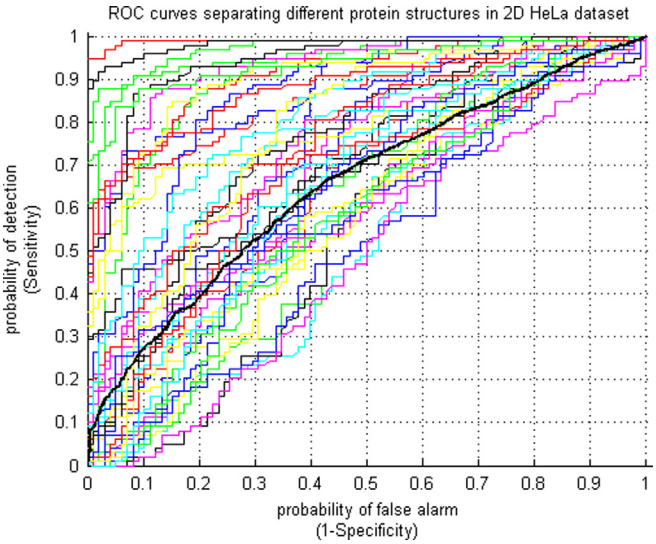


Fig. 8. ROC curve using uniform rotation invariant LTPs on radius 3 for 2D HeLa dataset. AUC value is 0.66.

Table 2
Performance of poly-SVM using LTPs for balanced mRMR based HeLa dataset.

m	*R*	*N*	*θ*	*D*	Polynomial kernel				
					Acc	**Sen**	**Spe**	**MCC**	***F*-Score**
Rotation invariant	1	8	40	60	89.8	91.4	89.7	0.62	0.64
Uniform	1	8	40	89	93.3	94.0	93.2	0.72	0.73
Uniform rotation invariant	1	8	40	18	91.6	92.5	91.5	0.67	0.68
Uniform	2	16	80	365	95.2	95.9	95.1	0.78	0.80
Uniform rotation invariant	2	16	80	27	94.1	95.6	94.0	0.75	0.76
Uniform rotation invariant	3	24	80	45	**95.1**	95.7	95.0	0.78	0.79

between the protein structures in 2D HeLa dataset. The features maintained discrimination between the classes as is evident from the *ROC* curve illustrated in Fig. 8, which is based on one feature rather than on the whole feature space.

Since we have 10 classes, there are 45 possible pairwise combinations and consequently, there are 45 *ROC* curves. Therefore, we calculated the average of all the curves. The good performance shown by LTPs demonstrated that the information extracted through various mappings by varying the radius of observation is useful for protein image classification of 2D HeLa images.

From the experimental results, it is observed that effective feature extraction strategy coupled with data balancing technique enable the classifier to make good predictions. Due to data balancing, all the classes have equal representation during the learning phase. Hence, a simple model is capable of predicting subcellular locations with quite high accuracy.

For all the LTPs types, we have tested the performance of SVM by reducing their dimensions using mRMR to different sizes as shown in Table 2.

Any significant improvement has not been observed in the performance of proposed prediction model. This revealed that patterns extracted using LTPs have as much strength as might be produced by using any other feature extraction strategy by employing mRMR.

From Tables 1 and 2, we observed that Uniform LTPs in 8-pixels neighborhood with mRMR showed slight improvement over LTPs without mRMR, rest mRMR did not add any strength to the discriminative power of LTPs. We have also tested the performance

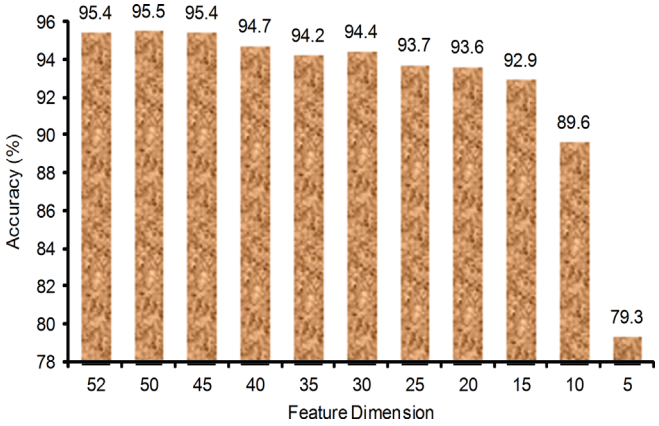


Fig. 9. mRMR based results using uniform rotation invariant LTPs on radius 3 for 2D HeLa dataset.

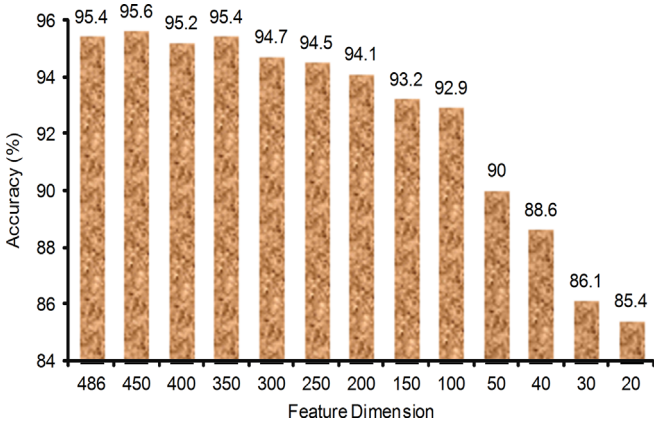


Fig. 10. mRMR based results using uniform LTPs on radius 2 for 2D HeLa dataset.

of SVM for various dimensionalities of the same features, which could be found in Supplementary Tables 1–3.

We have also tested the performance of mRMR on the two highest performing features spaces of LTPs shown in Table 1, in order to obtain the in depth analysis of the performance of these features. We have gradually decreased the features dimensions and observed the performance of SVM on each feature subspace. From Fig. 9, we observed that SVM performed well and gained the highest accuracy of 95.5% using the 50-D feature subspace. However, this 0.1% improvement cannot be claimed a significant achievement. Similar conclusion can be drawn for Fig. 10 where the improvement is merely 0.2%.

It is inferred that these features are much informative and there is not any severe overlapping in the information. As we further decreased the dimension of the feature space, the accuracy also got decreased, resulted in the lowest accuracy of 79.3% using 5-D feature subspace.

4.2. Prediction performance for CHO dataset

Table 3 demonstrates the prediction results of polynomial SVM of degree 2 using LTPs for CHO dataset. Uniform rotation invariant LTPs features on radius and neighborhood of 3 and 24 pixels, respectively, have achieved the highest prediction rates among other LTPs. Its accuracy is 90.7% with 0.65 MCC and 0.69 *F*-measure values. Hence, these features are observed to have better discrimination power compared to other LTPs. The performance of the proposed prediction system is also reported using the unbalanced CHO cell lines as shown in Supplementary Table 6.

Table 3
Performance of poly-SVM using LTPs for balanced CHO dataset.

m	*R*	*N*	*θ*	*D*	Polynomial kernel				
					Acc	**Sen**	**Spe**	**MCC**	***F*-Score**
Rotation invariant	1	8	30	72	70.4	59.6	71.9	0.22	0.33
Uniform	1	8	30	118	72.8	61.0	74.4	0.25	0.35
Uniform rotation invariant	1	8	30	20	68.4	55.7	70.2	0.18	0.30
Uniform	2	16	30	486	82.2	59.8	50.9	0.08	0.31
Uniform rotation invariant	2	16	30	36	83.9	77.1	84.9	0.48	0.54
Uniform rotation invariant	3	24	30	**52**	**90.7**	85.3	91.5	0.65	0.69

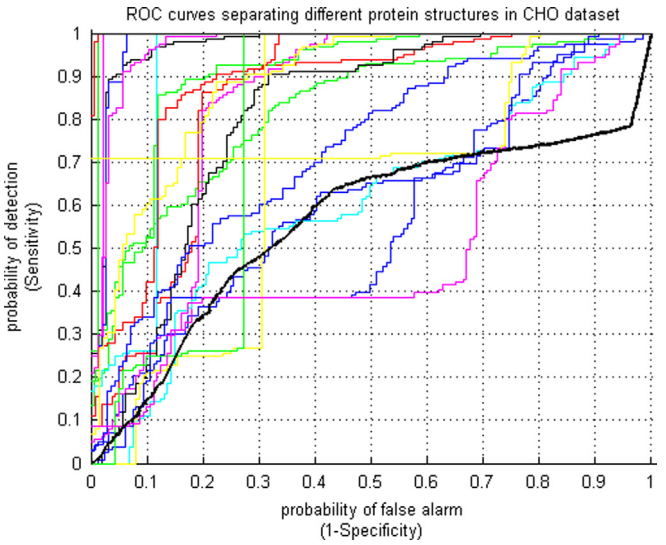


Fig. 11. ROC curve using uniform rotation invariant LTPs on radius 3 for CHO dataset. AUC value is 0.56.

Table 4
Performance of poly-SVM using LTPs for balanced mRMR based CHO dataset.

m	*R*	*N*	*θ*	*D*	Polynomial kernel				
					Acc	***Sen***	***Spe***	***MCC***	***F*-Score**
Rotation invariant	1	8	30	36′	68.9	57.2	70.5	0.19	0.31
Uniform	1	8	30	59	70.3	57.6	72.1	0.21	0.32
Uniform rotation invariant	1	8	30	10	64.2	52.0	65.9	0.12	0.26
Uniform	2	16	30	243	80.4	67.3	82.3	0.38	0.45
Uniform rotation invariant	2	16	30	18	80.7	71.8	81.9	0.41	0.47
Uniform rotation invariant	3	24	30	26	89.3	83.6	90.1	0.61	0.65

The highest accuracy using SMOTE is 8.6% higher than the accuracy without using SMOTE for the same LTPs features. This revealed that SMOTE has an effective role in the performance of the proposed prediction system.

The AUC is also computed for balanced LTPs features as shown in Fig. 11. The value of AUC obtained is 0.86. The black thick line shows the average of 28 ROC curves, which are produced from pair wise combinations of 8 classes.

We have also tested the performance of SVM using the LTPs patterns with mRMR. The performance of the prediction system is degraded in the presence of mRMR as shown in Table 4.

4.3. Ensemble generation and analysis

Ensemble is generated using the majority voting scheme. As is evident from Tables 1–4, various mappings have been used in LTPs computation; we have assigned serial numbers to different

Table 5
The serial numbers, attached to the mapping used in LTPs computation, representing a particular LTPs feature in Tables 6 and 7.

S. no.	*m*	*R*	*N*	*D*
1	Uniform	1	8	118
2	Rotation invariant	1	8	72
3	Uniform rotation invariant	1	8	20
4	Uniform rotation invariant	2	16	36
5	Uniform	2	16	486
6	Uniform rotation invariant	3	24	**52**

Table 6
Ensemble performance of different combinations of LTPs based poly-SVM classifications for balanced HeLa dataset.

S. no.	Ensemble participants						Polynomial kernel					
	1	**2**	**3**	**4**	**5**	**6**	**Acc**	**Sen**	**Spe**	**MCC**	***F*-Score**	***Q*-value**
1	√	√	√	√	√	–	99.8	99.9	99.8	0.99	0.99	0.20
2	√	√	√	√	–	√	99.8	99.9	99.8	0.99	0.99	0.07
3	√	√	√	–	√	√	99.8	99.9	99.8	0.99	0.99	0.08
4	√	√	–	√	√	√	99.8	99.9	99.8	0.99	0.99	0.11
5	√	–	√	√	√	√	**100**	100	100	1	1	0.01
6	–	√	√	√	√	√	99.7	99.9	99.7	0.98	0.98	0.12

Table 7
Ensemble performance of different combinations of LTPs based poly-SVM classifications for balanced CHO dataset.

S. no.	Ensemble participants						Polynomial kernel					
	1	**2**	**3**	**4**	**5**	**6**	**Acc**	**Sen**	**Spe**	**MCC**	***F*-Score**	***Q*-value**
1	√	√	√	√	√	–	91.7	87.3	92.4	0.69	0.72	0.38
2	√	√	√	√	–	√	92.8	89.1	93.4	0.72	0.75	0.36
3	√	√	√	–	√	√	92.9	88.8	93.5	0.72	0.75	0.36
4	√	√	–	√	√	√	94.7	91.4	95.1	0.78	0.81	0.34
5	√	–	√	√	√	√	**95.0**	91.6	95.4	0.79	0.81	0.33
6	–	√	√	√	√	√	93.6	90.2	94.1	0.75	0.77	0.38

mappings as shown in Table 5. We have chosen 5 SVM classifications among 6 for the ensemble generation. The reason is to have odd number of voters so that ties are avoided in decision making. The ensemble accuracies are presented in Tables 6 and 7 for 2D HeLa and CHO datasets, respectively.

In Table 6, the ensemble accuracy, sensitivity, specificity, MCC, *F*-measure, and average *Q*-value for 2D HeLa dataset are presented. The highest accuracy 100% is achieved for the 5th group. The average *Q*-value 0.01 shows that highest diversity exists among the voters of this group. The *Q*-value for the generated ensemble shows the similarity among the classifiers, hence, minimum is this value maximum is the diversity. The diversity values are obtained by subtracting this value from 1 and are presented in Supplementary Table 7.

Similarly, in Table 7, the highest ensemble accuracy achieved by the 5th group for CHO dataset is 95%, which is quite satisfactory. The diversity among the voters of this group also revealed that the effectiveness of this group in predicting subcellular localization of proteins from fluorescence microscopy images is good. The diversity values are provided in Supplementary Table 8.

5. Comparison with existing approaches

We have compared our proposed approach with many state-of-the-art existing approaches in the literature as presented in Table 8.

Table 8
Performance comparison with other approaches.

Method	Details of the used techniques	Accuracy	
		HeLa	**CHO**
Chebira et al. (2007)	Multi-resolution subspaces, weighted majority voting	95.4	–
Nanni and Lumini (2008)	Random subspace ensemble of NNs	94.2	–
Nanni et al. (2010c)	Random subspace ensemble of NNs, AdaBoost ensemble of weak learners, sum rule	97.5	–
(Nanni et al. (2010a)	Random subspace ensemble of NNs	95.8	–
Nanni et al. (2010b)	SVM, random subset of features, 50 classifiers, sum rule	93.2	–
Lin et al. (2007)	Variant of AdaBoost named as AdaBoost.ERC	93.6	94.7
Proposed approach	Majority voting based Ensemble	**100**	**95.0**

Accuracy of 95.4% is achieved by the proposed technique in Chebira et al. (2007). Nanni and Lumini reported accuracies of 94.2% for the 2D HeLa dataset (Nanni and Lumini, 2008). In another approach, Nanni et al. (2010c) have achieved the highest accuracy of 97.5% for the same dataset. Nanni et al. (2010a) also obtained 95.8% accuracy using this dataset. Nanni et al. (2010b) have reported the accuracy of 93.2% for 2D HeLa dataset Lin et al. (2007) have reported 93.6% and 94.7% accuracies, respectively, for HeLa and CHO datasets .

From Table 8, we can observe that our proposed approach outperforms all the existing approaches. The performance accuracy is 2.5% and 0.3% higher for the HeLa and CHO datasets, respectively, than the highest accuracies of the existing approaches.

6. Conclusion

In this paper, we have proposed an efficient ensemble approach that is based on LTPs based classifications, balanced feature space using SMOTE and polynomial SVM. It has been shown that the proposed technique performed better than many state-of-the-art approaches reported in the literature. We have shown that the employed combination of LTPs, SMOTE, and SVM is capable of producing comparable prediction outcomes. Uniform rotation invariant LTPs achieved the highest success rates with reasonably reduced feature space. The highest success rate obtained by the proposed model is 95.4% with 52-D feature space among the individual classifiers. It is further observed that the discriminative power of LTPs has not been improved with mRMR based feature selection.

We showed that the performance of the proposed model is outstanding because LTPs operator is less sensitive to noise in homogenous image regions; hence, having more discrimination power compared to other texture based operators. Further, LTPs used different mapping techniques to extract features from an image, this lead to extracting different information from the same image. In addition, application of SMOTE data balancing technique makes it possible for the classifier to have equal samples for training from all the classes. It is observed from the simulation results that the performance of SMOTE increases with the increase in the imbalance, present in the dataset. As can be seen, in case of 2D HeLa dataset, the improvement is very little after balancing the data. On the other hand, in case of CHO dataset, the enhancement is more because this dataset possesses more unbalanced nature compared to that of HeLa dataset.

In conclusion, SVM takes full advantage of exploiting the discrimination power of the adopted feature extraction strategy and balanced feature space. Further, their ensemble produced outstanding results. The experimental results confirmed the significance of balanced feature space exploration and efficient exploitation by SVM. In addition, it is validated that the prediction systems may be developed with perfect prediction rate.

Acknowledgment

This work is supported by the Higher Education Commission of Pakistan under the indigenous PhD scholarship program 17-5-4 (Ps4-124)/HEC/Sch/2008).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2013.08.017.

References

Boland, M.V., Murphy, R.F., 2001. 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.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357.
Chebira, A., Barbotin, Y., Jackson, C., Merryman, T., Srinivasa, G., Murphy, R.F., Kovacevic, J., 2007. A multiresolution approach to automated classification of protein subcellular location images. BMC Bioinformatics 8, 210.
Chen, W., Feng, P.M., Lin, H., Chou, K.-C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Research 41, e68.
Chen, X., Velliste, M., Murphy, R.F., 2006. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. Cytometry Part A, Journal of the International Society for Advancement of Cytometry 69A, 631–640.
Chen, Y.-K., Li, K.-B., 2013. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. Journal of Theoretical Biology 318, 1–12.
Chou, K.-C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol 44, 60) 43, 246-255.
Chou, K.-C., 2011. some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273, 236–247.
Chou, K.-C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Molecular Biosystems 9, 1092–1100.
Chou, K.-C., Wu, Z.-C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.
Chou, K.-C., Wu, Z.-C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Molecular Biosystems 8, 629–641.
Esmaeili, M., Mohabatkar, H., Mohsenzadeh, S., 2010. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. Journal of Theoretical Biology 263, 203–209.
Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., Torres, A., 2009. Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. Journal of Theoretical Biology 257, 17–26.
Gunn, S.R., 1998. Support Vector Machines for Classification and Regression, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science. University of Southampton, Southampton.
Hamilton, N.A., Pantelic, R.S., Hanson, K., Teasdale, R.D., 2007. Fast automated cell phenotype image classification. BMC Bioinformatics 8, 110.
Hand, D.J., Till, R.J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning 45, 171–186.
Hayat, M., Khan, A., 2011. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. Journal of Theoretical Biology 271, 10–17.
Hayat, M., Khan, A., 2012. MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM. Journal of Theoretical Biology 292, 93–102.
Hayat, M., Khan, A., Yeasin, M., 2011. Prediction of membrane proteins using split amino acid composition and ensemble classification. Amino Acids 42, 2447–2460.
He, Z.-S., Shi, X.-H., Kong, X.-Y., Zhu, Y.-B., Chou, K.-C., 2012. A novel sequence-based method for phosphorylation site prediction with feature selection and analysis. Protein & Peptide Letters 19, 70–78.
He, Z., Zhang, J., Shi, X.-H., Hu, L.-L., Kong, X., Cai, Y.-D., Chou, K.-C., 2010. Predicting drug-target interaction networks based on functional groups and biological features. PLoS One 5, e9603.
Hu, Y., Murphy, R.F., 2004. Automated interpretation of subcellular patterns from immunofluorescence microscopy. Journal of Immunological Methods 290, 93–105.
Huang, T., Chen, L., Cai, Y.-D., Chou, K.-C., 2011. Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property. PLoS One 6, e25297.
Huang, T., He, Z.-S., Cui, W.R., Cai, Y.-D., Shi, X.H., Hu, L.-L., Chou, K.-C., 2013. A sequence-based approach for predicting protein disordered regions. Protein & Peptide Letters 20, 243–248.
Huang, T., Shi, X.-H., Wang, P., He, Z., Feng, K.-Y., Hu, L., Kong, X., Li, Y.-X., Cai, Y.-D., Chou, K.-C., 2010. Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS One 5, e10972.
Khan, A., Khan, M.F., Choi, T.-S., 2008. Proximity based GPCRs prediction in transform domain. Biochemical and Biophysical Research Communications 371, 411–415.
Khan, A., Majid, A., Hayat, M., 2011. CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of Pseudo amino acid composition. Computational Biology and Chemistry 35, 218–229.
Khosravian, M., FK, F.K.F., Beigi, M.M., M, M.B., Mohabatkar, H., 2013. Predicting antibacterial peptides by the concept of Chou's Pseudo-amino acid composition and machine learning methods. Protein & Peptide Letters 20, 180–186.
Li, B.-Q., Hu, L.-L., Niu, S., Cai, Y.-D., Chou, K.-C., 2012a. Predict and analyze *S*-nitrosylation modification sites with the mRMR and IFS approaches. Journal of Proteomics 75, 1654–1665.
Li, B.-Q., Huang, T., Liu, L., Cai, Y.-D., Chou, K.-C., 2012b. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS One 7, e33393.
Li, B.-Q., Hu, L.-L., Chen, L., Feng, K.-Y., Cai, Y.-D., Chou, K.-C., 2012c. Prediction of protein domain with mRMR feature selection and analysis. PLoS One 7, e39308.
Li, S., Kwok, J.T., Zhu, H., Wang, Y., 2003. Texture classification using the support vector machines. Pattern Recognition 36, 2883–2893.
Lin, C.-C., Tsai, Y.-S., Lin, Y.-S., Chiu, T.-Y., Hsiung, C.-C., Lee, M.-I., Simpson, J.C., Hsu, C.-N., 2007. Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization. Bioinformatics 23, 3374–3381.
Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.-C., 2013. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Molecular Biosystems 9, 634–644.
Mei, S., 2012. Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. Journal of Theoretical Biology 293, 121–130.
Meynet, J., Thiran, J.-P., 2010. Information theoretic combination of pattern classifiers. Pattern Recognition 43, 3412–3421.
Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein & Peptide Letters 17, 1207–1214.
Mohabatkar, H., Beigi, M.M., Esmaeili, A., 2011. Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology 281, 18–23.
Murphy, R.F., 2004. Automated interpreation of subcellular location patterns. IEEE International Symposium on Biomedical Imaging: Nano to Macro 1, 53–56.
Murphy, R.F., Boland, M.V., Velliste, M., Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images, In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, La Jolla/ San Diego, CA, USA 2000, pp. 251–259.
Murphy, R.F., Velliste, M., Porreca, G., 2003. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. Journal of VLSI Signal Processing 35, 311–321.
Nanni, L., Lumini, A., 2008. A reliable method for cell phenotype image classification. Artificial Intelligence in Medicine 43, 87–97.
Nanni, L., Brahnam, S., Lumini, A., 2010a. Novel features for automated cell phenotype image classification. Advances in Computational Biology: Advances in Experimental Medicine and Biology (AEMB) 680, 207–213.
Nanni, L., Brahnam, S., Lumini, A., Selecting the best performing rotation invariant patterns in local binary/ternary patterns. In: Proceedings of the International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV'10), Las Vegas, Nevada, USA 2010b, pp. 369–375.
Nanni, L., Lumini, A., Lin, Y.-S., Hsu, C.-N., Lin, C.-C., 2010c. Fusion of systems for automated cell phenotype image classification. Expert Systems with Applications 37, 1556–1562.
Sahu, S.S., Panda, G., 2010. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational Biology and Chemistry 34, 320–327.
Shi, J.-Y., Zhang, S.-W., Pan, Q., Cheng, Y.-M., Xie, J., 2007. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33, 69–74.
Srinivasa, G., Merryman, T., Chebira, A., Kovacevic, J., Mintos, A., 2006. Adaptive multiresolution techniques for subcellular protein location classification. IEEE International Conference on Acoustics, Speech and Signal Processing 5, 14–19.

Tahir, M., Khan, A., Majid, A., 2012. Protein subcellular localization of fluorescence imagery using spatial and transform domain features. Bioinformatics 28, 91–97.
Tscherepanow, M., Jensen, N., Kummert, F., 2008. An incremental approach to automated protein localisation. BMC Bioinformatics 9, 445.
Wu, Z.C., Xiao, X., Chou, K.-C., 2011. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. Molecular Biosystems 7, 3287–3297.
Wu, Z.C., Xiao, X., Chou, K.-C., 2012. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. Protein & Peptide Letters 19, 4–14.
Xiao, X., Wu, Z.-C., Chou, K.-C., 2011a. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 6, e20592.
Xiao, X., Wu, Z.C., Chou, K.-C., 2011b. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. Journal of Theoretical Biology 284, 42–51.
Xu, Y., Ding, J., Wu, L.Y., Chou, K.-C., 2013. iSNO-PseAAC: predict cysteine *S*-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8, e55844.
Yi, X.-F., He, Z.-S., Chou, K.-C., Kong, X.-Y., 2012. Nucleosome positioning based on the sequence word composition. Protein & Peptide Letters 19, 79–90.
Zhang, L., Liao, B., Li, D., Zhu, W., 2009. A novel representation for apoptosis protein subcellular localization prediction using support vector machine. Journal of Theoretical Biology 259, 361–365.
Zhang, S.-W., Zhang, Y.-L., Yang, H.-F., Zhao, C.-H., Pan, Q., 2008. Using the concept of Chou′s pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34, 565–572.
Zhang, T.-L., Ding, Y.-S., 2007. Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33, 623–629.