BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 1 2012, pages 91–97
doi:10.1093/bioinformatics/btr624

Systems biology

Advance Access publication November 15, 2011

Protein subcellular localization of fluorescence imagery using spatial and transform domain features

Muhammad Tahir, Asifullah Khan* and Abdul Majid

Department of Computer and Information Sciences, PIEAS, Islamabad, Pakistan

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Subcellular localization of proteins is one of the most significant characteristics of living cells. Prediction of protein subcellular locations is crucial to the understanding of various protein functions. Therefore, an accurate, computationally efficient and reliable prediction system is required.

Results: In this article, the predictions of various Support Vector Machine (*SVM*) models have been combined through majority voting. The proposed ensemble *SVM-SubLoc* has achieved the highest success rates of 99.7% using hybrid features of Haralick textures and local binary patterns (*HarLBP*), 99.4% using hybrid features of Haralick textures and Local Ternary Patterns (*HarLTP*). In addition, *SVM-SubLoc* has yielded 99.0% accuracy using only local ternary patterns (*LTPs*) based features. The dimensionality of *HarLBP* feature vector is 581 compared with 78 and 52 for *HarLTP* and *LTPs*, respectively. Hence, *SVM-SubLoc* in conjunction with *LTPs* is fast, sufficiently accurate and simple predictive system. The proposed *SVM-SubLoc* approach thus provides superior prediction performance using the reduced feature space compared with existing approaches.

Availability: A web server accompanying the proposed prediction scheme is available at http://111.68.99.218/ *SVM-SubLoc*

Contact: asif@pieas.edu.pk; khan.asifullah@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 28, 2011; revised on October 11, 2011; accepted on November 7, 2011

1 INTRODUCTION

Comprehension of the functions of proteins is of prime importance in the field of biological sciences (Nanni *et al.*, 2010c). One of the significant characteristics of proteins is its subcellular localization that reveals precious information regarding the working of proteins (Murphy *et al.*, 2000). Determining protein subcellular locations is significant to the understanding of various protein functions. For instance, during the drug discovery process, precise knowledge of the subcellular localization of proteins can be useful for the identification of drugs. In addition, the effectiveness of drugs can be estimated by knowing the exact locations of proteins before and after using the drugs (Khan *et al.*, 2011; Srinivasa *et al.*, 2006).

Fluorescence microscopy is frequently used to determine subcellular localization of proteins in cells. Images of protein locations are normally analyzed in traditional ways, which are time consuming and prone to errors (Chebira *et al.*, 2007; Nanni *et al.*, 2010c). Automated approaches are thus required for the classification of such images. These methods need to be computationally efficient and accurate. Fortunately, considerable progress is made in recent years for the development of computational methods that can automatically determine the subcellular protein locations from fluorescence microscopy images (Lin *et al.*, 2007; Murphy *et al.*, 2002, 2003). Murphy *et al.* have developed subcellular location feature (*SLF*) sets and trained a back propagation neural network (*BPNN*) to test the performance of these features using 2*D HeLa* dataset (Murphy *et al.*, 2000). Boland and Murphy have tested the performance of Haralick textures, Zernike moments, *SLF1* and different combination of these three by employing the *BPNNs* (Boland and Murphy, 2001). Murphy *et al.* (2003) have trained *BPNN* using enhanced feature sets consisting of Haralick textures, Zernike moments and morphological features. Hamilton *et al.* (2007) have employed Support Vector Machine (*SVM*) using various feature extraction strategies including threshold adjacency statistics (*TASs*), Zernike moments as well as a hybrid of *TASs* and Haralick textures. Chebira *et al.* (2007) have developed automated classification system for protein subcellular location images in multiresolution subspaces. They have employed *ANN* at different decomposition levels to obtain the classification results, which are then combined through weight assignment. Nanni *et al.* (2009) have utilized random subspace of Levenberg–Marquardt neural networks and AdaBoost learning algorithm. In addition, they have employed the fusion between these two ensemble classifiers, while different local and global descriptors have been implemented as feature sets. Recently, Nanni *et al.* (2010a) have employed a random subspace of Levenberg–Marquardt neural networks using optimized sets of various feature extraction strategies including Wavelet features, Haralick textures, local binary patterns (*LBPs*), local ternary patterns (*LTPs*) and TASs for 2*D HeLa* and *LOCATE* mouse protein datasets.

The models proposed by different researchers have still margin in improving the prediction accuracy and reducing the dimensionality of the feature space. The aim of this study is to develop an accurate and simple system compared with the existing approaches. We thus develop both individual and hybrid feature based classification approaches for the prediction of protein subcellular localization. In the proposed approach, we employ Haralick textures, Zernike moments, *LBPs*, *LTPs* and *TASs* based feature extraction strategies. Different hybrid feature sets are formed by concatenating these features. The performance of various kernels of *SVM* has been investigated using these features. To enhance the performance of

*To whom correspondence should be addressed.

the proposed model, the success rates of different *SVMs* have been combined through majority voting. Discrete Wavelet Transforms is used for the extraction of Haralick textures and Zernike moments only. For this purpose, we have decomposed the image upto four levels so that the best decomposition level could be detected. Then, we have used statistical measures to acquire Haralick textures and Zernike moments from each decomposed image.

Rest of the article is organized as follows: Datasets and different feature extraction strategies are described in Section 2. The proposed approach is presented in Section 3. Results and discussions are elaborated in Section 4. Conclusions are drawn at the end.

2 METHODS

2.1 Datasets

Three datasets have been used to evaluate the performance of our proposed scheme including *2D HeLa*, LOCATE Endogenous and LOCATE Transfected datasets. The *2D HeLa* dataset contains 862 single-cell images, each of size 382×382, distributed in 10 classes (Chebira *et al.*, 2007). LOCATE Endogenous and LOCATE Transfected datasets contain 502 and 553 images, respectively. Each image is of size 768×512, having up to 13 cells. LOCATE Endogenous and LOCATE Transfected images are distributed in 10 and 11 classes, respectively (Nanni *et al.*, 2010a) Classes and images per each class of *2D HeLa* and LOCATE datasets are provided in Supplementary Tables S12 and S13, respectively.

2.2 Feature extraction strategies

In this work, we have employed various texture based feature extraction strategies such as Haralick textures, Zernike moments, *LBPs*, *LTPs* and *TASs*. We describe these feature extraction strategies as follows.

2.2.1 Haralick texture features Haralick features are texture based statistical measures utilized by a number of researchers for classification (Hamilton *et al.*, 2006; Haralick, 1979; Nanni *et al.*, 2010c). A Spatial Gray Level Dependence Matrix (*SGLD*) of size $N\times N$ is first obtained for an image with N gray levels at certain angle θ (i.e. 0°, 45°, 90° and 135°) at some distance d where d is measured in terms of pixel distance. In this work, d is set to one. Then, 13 statistical measures are calculated from *SGLD* matrix, namely energy, correlation, inertia, entropy, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy and two information measures of correlation. In this work, features along horizontal and vertical directions are combined by averaging. Similarly, features along diagonal and off-diagonal directions are also combined by averaging. As a result, we have obtained a total of 26 features for each image without Discrete Wavelet Transform (*DWT*). With *DWT*, the number of images at each decomposition level varies as well as the dimension of feature space as shown in Table 1.

2.2.2 Zernike moment-based features A set of complex polynomials, which form a complete orthogonal set over the interior of the unit circle, provides the basis to compute Zernike moments. Zernike moments do not bear redundant information between the moments because Zernike polynomials are orthogonal to each other. They are computationally inexpensive compared with other texture based features (Hu and Murphy, 2004). Zernike moments of an image can be calculated using Equation (1) (Boland *et al.*, 1998).

$$Z_{nl}=\frac{n+1}{\pi}\sum_{x,y}V^{*}_{nl}(x,y)I(x,y) \tag{1}$$

where $I(x,y)$ represents the pixel intensity at position $(x,y), x^2+y^2\le1, 0\le l\le n, n-l$ is even. V^{*}_{nl} indicates complex conjugate of the Zernike polynomial

Table 1. Performance of SVM-SubLoc using Haralick texture features with/without DWT

		lin	*poly*	*RBF*	*sig*	Ensemble			
L	*D*	Acc				Acc	*MCC*	*F*-score	*Q*-statistic
0	26	75.6	75.9	76.4	69.6	87.7	0.59	0.60	0.29
1	104	80.5	80.9	81.7	80.3	92.5	0.71	0.72	**0.18**
2	416	81.9	83.2	**84.1**	79.8	**93.2**	**0.73**	**0.74**	0.20
3	1664	77.3	78.3	80.6	77.9	90.2	0.65	0.66	0.30
4	6656	79.5	80.6	80.7	79.5	92.6	0.72	0.73	0.23

$L=0$ represent without DWT. Highest values obtained are represented in bold.

Table 2. Performance of *SVM-SubLoc* using Zernike moment-based features with/without *DWT*

		lin	*poly*	*RBF*	*sig*	Ensemble			
L	*D*	Acc				Acc	*MCC*	*F*-score	*Q*-statistic
0	49	28.3	43.8	46.5	15.5	46.7	−0.02	0.16	0.02
1	196	35.9	50.3	52.2	24.4	57.8	0.12	0.23	0.09
2	784	39.0	53.4	56.2	15.1	**58.5**	**0.15**	**0.25**	**0.01**
3	3136	50.2	52.0	60.5	11.3	54.5	0.10	0.22	0.24
4	12 544	44.0	48.0	**67.8**	11.0	56.7	0.14	0.24	0.11

$L=0$ represent without DWT. Highest values obtained are represented in bold.

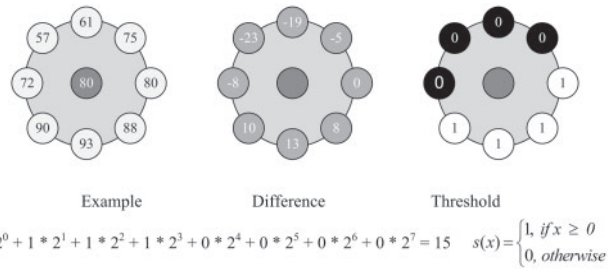

Fig. 1. Procedure of *LBPs* code generation.

of degree n and angular dependence l. Its value can be computed as follows:

$$V_{nl}(x,y)=\sum_{m=0}^{(n-l)/2}\frac{(-1)^m(x^2+y^2)^{n/2-m}e^{il\theta}(n-m)!}{m!(\frac{n-2m+l}{2})!(\frac{n-2m-l}{2})!} \tag{2}$$

where $0\le l\le n, n-l$ is even and $\theta=\tan^{-1}(y/x)$.

In this work, we have obtained Zernike moments of order 12 as used by Boland *et al.* (1998); Chebira *et al.* (2007); Hamilton *et al.* (2007). These features are extracted in spatial and transform domains. We have employed *DWT* for the transformation, which decomposes each image into four subimages. We have computed features at each decomposition level separately. Due to the varying number of images at each decomposition level, dimensions of feature vectors also vary as shown in Table 2.

2.2.3 Local Binary Patterns LBPs is a texture-based feature extraction strategy for gray level patterns in an image, proposed by Ojala *et al.* (1996). *LBPs* operator evaluates the binary differences between the gray value of the central pixel c and the gray values of P pixels in the neighborhood on a circle of radius R around c. (Nanni *et al.*, 2010a, b; Nanni and Lumini, 2008). The procedure of obtaining *LBPs* code is depicted in Figure 1.

$s(x)$ shows the value of each pixel p after applying the threshold. The *LBPs* code is generated according to Equation (3):

$$\mathrm{LBP}_{P,R}=\sum_{p=0}^{P-1}s(g_p-g_c)2^p \tag{3}$$

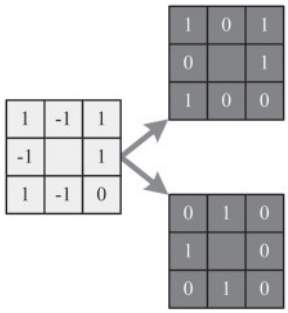

Fig. 2. *LTP* code splits into two *LBP* codes.

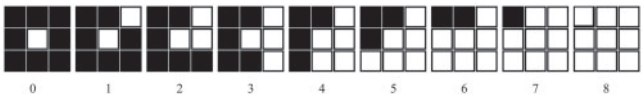

Fig. 3. *TAS*: 0–8 neighboring white pixels of a central white pixel in a 3×3 neighborhood.

Where g_c and g_p represent the gray-level values of central and neighboring pixels, respectively. *LBPs* are computed using uniform ($u2$), rotation invariant (ri) and uniform rotation invariant ($riu2$) mappings on three different configurations: ($R=1$, $N=8$), ($R=2$, $N=16$) and ($R=3$, $N=24$) where R and N indicate radius and neighborhood, respectively.

2.2.4 Local Ternary Patterns *LTPs* is based on the generalization of *LBPs* (Tan and Triggs, 2007). In *LTPs*, the difference between a central pixel c and its neighbor u is based upon a ternary value according to a threshold θ as given by Equation (4).

$$s(u) = \begin{cases} 1 & \text{if } u \geq c+\theta \\ -1 & \text{if } u \leq c-\theta \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

To reduce computational complexity, the ternary pattern is split into two binary patterns according to its positive and negative components, as shown in Figure 2. The histograms computed from the component binary patterns are concatenated to obtain the feature vector for the ternary pattern (Nanni *et al.*, 2010a, b).

LTPs are also computed using uniform ($u2$), rotation invariant (ri) and uniform rotation invariant ($riu2$) mappings on three different configurations i.e. ($R=1, N=8$), ($R=2, N=16$) and ($R=3, N=24$) where R and N indicate value of radius and number of neighborhood pixels, respectively.

2.2.5 Threshold Adjacency Statistics *TASs* based features are computationally inexpensive and efficient metric for classifying subcellular localization images (Hamilton *et al.*, 2007). *TASs* features are computed from three 9-bin histograms, obtained from three different binary images, generated using three different thresholds (Nanni and Lumini, 2008). These features are calculated as follows. First, a threshold is applied to the image to produce a binary image. Then, nine statistics are computed from that binary image as shown in Figure 3.

The first statistic is the number of white pixels that have no white neighbors, the second statistic is the number of white pixels that have exactly one white neighbor and the third statistic is the number of white pixels that have exactly two white neighbors. This process is repeated up to eight white neighboring pixels for 8-bit gray scale image. Two other sets of *TASs* are calculated in similar fashion. Each set of *TAS* is computed for binary images with pixel intensities in the range of μ to 255, $\mu-\theta$ to 255 and $\mu+\theta$ to 255

Table 3. Performance of *SVM-SubLoc* using *TAS*

		lin	*poly*	*RBF*	*sig*	Ensemble			
θ	D	Acc				Acc	*MCC*	*F*-score	*Q*-statistic
140	27	77.6	80.3	**81.0**	70.7	**91.6**	**0.69**	**0.70**	**0.31**

Highest values obtained are represented in bold.

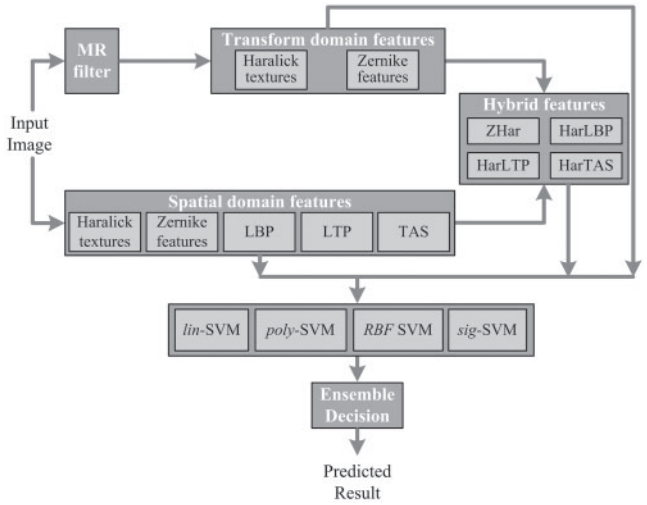

Fig. 4. Framework of the proposed *SVM-SubLoc*.

where μ is the average pixel intensity and θ is the user defined threshold. The value of threshold θ should not be <30 since less values are considered as background intensities (Hamilton *et al.*, 2007). We have observed the best results when $\theta=140$ as shown in Table 3.

2.2.6 Hybrid features A hybrid feature model is produced by combining different individual features to enhance the discrimination power of the feature space. The hybrid features include *ZHar* (Zernike + Haralick), *HarLBP* (Haralick + *LBPs*), *HarLTP* (Haralick + *LTPs*) and *HarTAS* (Haralick + *TASs*). Only *ZHar* hybrid features are formed both in spatial and transform domains.

2.3 SVM

SVM is a popular machine learning technique used in the field of pattern recognition and classification. The theoretical detail of *SVM* is available in the machine learning literature (Gunn, 1998; Hayat and Khan, 2011; Majid *et al.*, 2006). It was developed for binary classification problems. However, to employ *SVM* for multiclassification problems, a straightforward approach is to reduce the multiclassification to a series of binary classifications through one-versus-rest mechanism. For a k-class classification problem, k *SVMs* are built where the ith *SVM* is trained on every instance in the ith class with positive labels and all other instances with negative labels. In this study, we have used four different kernels; linear (*lin*), polynomial (*poly*) of degree 2, Radial Basis Function (*RBF*) and sigmoid (*sig*).

2.4 The proposed *SVM-SubLoc* approach

The framework of our proposed approach is shown in Figure 4. This figure highlights a new prediction model *SVM-SubLoc* based on different individual and hybrid feature sets. Individual features, such as Haralick textures and Zernike moments are extracted in transform and spatial domains using *DWT*. However, local binary patterns, local ternary patterns and TASs are obtained in spatial domain only. Hybrid features are constructed by concatenating these features in different combinations as described in Section 2.2.6. The

performance of various *SVMs* has been evaluated using these features. The predictions of these *SVMs* have been combined through majority voting to improve the performance of the proposed model. In case of a tie among the individual classifiers' voting, preference is given to the classifier with the highest performance.

3 RESULTS AND DISCUSSIONS

The Jackknife test is the most accurate and significantly efficient method for measuring the performance of algorithms (Hamilton *et al.*, 2007; Khan *et al.*, 2008). We have applied 5-fold cross-validation to explore the performance of the proposed approach. In case of *2D HeLa*, due to the slight imbalance, the input data is stratified before applying the cross-validation. The analysis of the results for *2D HeLa* dataset is presented in Section 3.1. The features that performed well on *2D HeLa* have also been tested on LOCATE Endogenous and Transfected datasets. In Section 3.2, only the best results on the two LOCATE datasets have been reported. However, detailed analysis using *LTPs*, *HarLBP* and *HarLTP* is provided in the Supplementary Material. We have used Accuracy, F-score, MCC and Q-statistic as performance measures.

$$Q_{\text{avg}} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k} \tag{5}$$

3.1 Performance analysis for *2D HeLa* dataset

We have computed the aforementioned measures using individual and hybrid features for *2D HeLa* dataset. In this section, first, we discuss the performance of *SVM-SubLoc* using individual feature sets on *2D HeLa* dataset. Afterwards, the performance of *SVM-SubLoc* will be reported for hybrid feature sets.

3.1.1 Performance of SVM-SubLoc using individual feature sets

In this section, we discuss our findings about the performance of *SVM-SubLoc* using individual features constructed in Section 2.2.

Performance of SVM-SubLoc using Haralick texture features

Table 1 shows the performance of *SVM-SubLoc* using Haralick textures with/without *DWT*. Here, L and D represent the decomposition level and dimension of the feature vector, respectively.

The individual *RBF-SVM* has achieved the highest success rates at all decomposition levels. Particularly, its highest accuracy 84.1% is observed at second decomposition level. Further, the *SVM-SubLoc* has brought a significant improvement in the prediction of subcellular protein locations for *2D HeLa* dataset. *SVM-SubLoc* has achieved the highest accuracy 93.2%, which is 9.1% higher than that of *RBF-SVM*. This shows that second level is the best decomposition level for this dataset. This is because at lower decomposition levels some valuable hidden information is lost. However, at higher decomposition levels, we obtained redundant information, which degraded the performance of the classifiers. The highest *MCC* value at second level shows that the prediction is quite encouraging. Similarly, the highest F-score value at this level indicates the best accuracy at this level. On the other hand, the highest diversity is obtained at first decomposition level with the Q value of 0.18.

Performance of SVM-SubLoc using Zernike moment-based features

The predictions of SVM-SubLoc using Zernike moments are reported in Table 2 with/without DWT. At 0th decomposition level, the highest yielded accuracy is 46.5%, which indicates poor feature extraction at that level. Among the base classifiers, *sig-SVM* has shown poor performance; only the accuracy of *RBF-SVM* is improved with the increase in the decomposition levels. We achieved the highest accuracy 67.8% at fourth level. Though, *SVM-SubLoc* has yielded the best accuracy 58.5% at second level due to the maximum diversity at this level as indicated by the Q value of 0.01. However, prediction quality and accuracy of the test are not reasonably good as shown by the *MCC* and F-score values at second decomposition level.

Table 4. Performance of *SVM-SubLoc* using *LBPs* for various mappings

				lin	*poly*	*RBF*	*sig*	Ensemble			
R	*N*	*m*	*D*	Acc				Acc	*MCC*	*F*-score	*Q*-statistic
1	8	u2	59	83.7	84.3	85.3	71.5	92.8	0.72	0.73	0.32
1	8	ri	36	83.5	82.4	82.7	72.9	92.8	0.72	0.73	0.34
1	8	riu2	10	81.5	82.5	82.5	74.9	92.5	0.71	0.72	0.33
2	16	u2	243	85.7	86.8	87.5	78.6	95.0	0.79	0.80	0.32
2	16	ri	4116	73.2	72.6	73.0	70.5	85.4	0.54	0.56	0.37
2	16	riu2	18	87.1	86.8	87.8	78.3	95.1	0.79	0.80	0.37
3	24	u2	555	85.0	86.3	87.4	78.8	**95.5**	**0.81**	**0.82**	**0.25**
3	24	riu2	26	85.7	86.4	**88.0**	80.5	94.8	0.78	0.79	0.34

Highest values obtained are represented in bold.

Performance of SVM-SubLoc using TASs

Table 3 demonstrates the predictions of SVM-SubLoc using TASs for gray images without DWT. In individual base classifiers, *RBF-SVM* has achieved the highest accuracy of 81.0% compared with other *SVMs*. The success rate of *SVM-SubLoc* is 91.6%, which is 10.6% higher than that of *RBF-SVM*. The increased ensemble accuracy shows the significance of the ensemble classifier. *MCC* value of 0.69 shows good quality of the prediction, whereas F-score value of 0.70 indicates fine accuracy of the performed test. The Q value of 0.31 reveals that results have 69% diversity. It has been analyzed during the experiments that *TASs* have produced the most significant results for $\theta = 140$. The discriminating capability of these features at this threshold is enhanced. The results achieved at other threshold values are presented in Supplementary Table S1 for comparison.

Performance of SVM-SubLoc using LBPs

The predicted outcomes of SVM-SubLoc using LBPs are presented in Table 4. Here, m represents mapping. In individual base learners, RBF-SVM has achieved the highest accuracy of 88.0% using riu2 LBPs for $R = 3$ and $N = 24$. However, SVM-SubLoc has yielded the highest accuracy of 95.5% using u2 LBPs for $R = 3$ and $N = 24$. The performance accuracy of SVM-SubLoc is 7.5% higher than that of RBF-SVM, which highlights the significance of the proposed ensemble technique. The maximum diversity of SVM-SubLoc has been observed when the features are extracted on the circle of radius 3 as is evident from Q value of 0.25. The best values of MCC (=0.81) and F-score (=0.82) are also obtained using u2 LBPs.

Performance of SVM-SubLoc using LTPs

In Table 5, the success rates of SVM-SubLoc using LTPs are shown for gray images without DWT. There is an additional parameter θ used by LTPs, which defines the threshold. In individual classifiers, poly-SVM has achieved the highest accuracy 94.4% for $R = 3$, $N = 24$, and $\theta = 80$ using riu2 LTPs. The highest accuracy of 99.0% achieved by SVM-SubLoc is 4.6% higher than that of poly-SVM. It has been investigated that LTPs have more discriminative capability

Table 5. Performance of SVM-SubLoc using LTPs for various mappings

					lin	*poly*	*RBF*	*sig*	Ensemble			
R	*N*	θ	*m*	*D*	Acc				Acc	*MCC*	*F*-score	*Q*-statistic
1	8	40	u2	118	90.4	90.4	90.8	78.1	97.7	0.89	0.90	0.35
1	8	40	ri	72	89.3	87.5	89.3	80.5	97.3	0.87	0.88	0.28
1	8	40	riu2	20	89.2	89.7	90.1	77.2	97.3	0.87	0.88	0.28
2	16	80	u2	486	92.8	93.0	92.9	84.1	98.2	0.91	0.92	0.21
2	16	80	riu2	36	91.8	92.9	93.5	85.1	98.6	0.93	0.93	0.26
3	24	80	riu2	52	93.8	**94.4**	93.8	86.3	**99.0**	**0.95**	**0.95**	**0.15**

Highest values obtained are represented in bold.

Table 6. Performance of *SVM-SubLoc* using *ZHar* with/without *DWT*

		lin	*poly*	*RBF*	*sig*	Ensemble			
L	*D*	Acc				Acc	*MCC*	*F*-score	*Q*-statistic
0	75	69.7	69.7	71.8	57.5	84.4	0.52	0.54	0.19
1	300	74.8	76.2	76.9	74.4	90.4	0.66	0.67	**0.12**
2	1200	80.0	80.5	**80.8**	77.2	**93.2**	**0.73**	**0.74**	0.24
3	4800	77.3	77.9	78.5	77.4	89.2	0.62	0.64	0.35
4	19 200	77.8	80.0	79.3	76.2	90.8	0.66	0.68	0.33

$L=0$ represent without DWT. Highest values obtained are represented in bold.

compared with other feature sets. As is evident from the Q value of 0.15, we observed the maximum diversity using riu2 LTPs for $R=3$ and $N=24$ and that is the main reason why the ensemble accuracy is high. An *MCC* value of 0.95 and *F*-score value of 0.95 also indicate that discrimination power *of riu2 LTPs* is better. It is evident from Table 5 that threshold values vary at $R=1$ and $R=2,3$. At smaller circles, the small value of θ performs well. However, at larger circles, the value of θ should be greater.

3.1.2 Performance of SVM-SubLoc using hybrid feature sets In this section, we discuss our findings regarding *SVM-SubLoc* using hybrid feature sets as given in Section 2.2.6.

Performance of SVM-SubLoc using ZHar

In Table 6, we present the predicted accuracies of *SVM-SubLoc* using the hybrid of Zernike and Haralick texture features with/without DWT. In individual classifiers, we have found the best performance of 80.8% for *RBF-SVM*.

The highest accuracy 93.2% obtained by *SVM-SubLoc* is 12.4% higher than that of *RBF-SVM*. The second decomposition level has been found to be the best level for discriminating subcellular location images. However, most diverse results are obtained at first level as indicated by *Q* value of 0.12. Highest values of *MCC* 0.73 and *F*-score 0.74 are also achieved at second level, which show that both quality of prediction and accuracy of the test are best at this level. The individual classifiers using these hybrid features do not produce better results compared with using their individual constituents. However, *SVM-SubLoc* has yielded the same accuracy as yielded by the ensemble using Haralick textures.

Performance of SVM-SubLoc using HarTAS

The success rates of *SVM-SubLoc* using *HarTAS* have been given is Table 7. Among different kernel-based *SVMs*, *poly-SVM* has achieved the highest accuracy of 87.9%, which is further enhanced by the ensemble *SVM-SubLoc* up to 96.2%.

This shows 8.3% improvement in the accuracy. The *Q* value of 0.32 indicates sufficient diversity among classifiers. Quality of the prediction and accuracy of the test are quite good as revealed by MCC value of 0.83 and F-score value of 0.84, respectively. The

Table 7. Performance of SVM-SubLoc using HarTAS

		lin	*poly*	*RBF*	*sig*	Ensemble			
θ	*D*	Acc				Acc	*MCC*	*F*-score	*Q*-statistic
140	53	87.0	**87.9**	86.0	83.0	**96.2**	**0.83**	**0.84**	0.32

Highest values obtained are represented in bold.

Table 8. Performance of SVM-SubLoc using HarLBP

				lin	*poly*	*RBF*	*sig*	Ensemble			
R	*N*	*m*	*D*	Acc				Acc	MCC	*F*-score	*Q*-statistic
1	8	u2	85	92.1	92.4	92.5	88.5	99.3	0.96	0.96	**0.14**
1	8	ri	62	89.2	89.9	89.0	84.9	97.0	0.86	0.87	0.31
1	8	riu2	36	88.3	90.2	89.6	87.9	97.5	0.88	0.89	0.28
2	16	u2	269	93.2	93.9	**94.4**	91.7	99.0	0.95	0.95	**0.14**
2	16	ri	4142	81.6	81.6	78.5	81.0	92.5	0.71	0.72	0.28
2	16	riu2	44	92.5	93.2	93.7	90.8	98.8	0.94	0.94	0.18
3	24	u2	581	93.2	93.6	**94.4**	90.9	**99.7**	**0.98**	**0.98**	0.20
3	24	riu2	52	91.8	92.6	92.8	89.7	99.3	0.96	0.96	0.34

Highest values obtained are represented in bold.

results at other threshold values are presented in Supplementary Table S2.

Performance of SVM-SubLoc using HarLBP

The results of SVM-SubLoc using hybrid feature set of HarLBP are shown in Table 8. In individual classifiers, RBF-SVM has achieved the highest accuracy of 94.4% for $R=3$ and $N=24$. The obtained accuracy is 18.0% higher than the highest accuracy using Haralick textures (at 0th level) as given in Table 1 and 6.4% higher than the highest accuracy using LBPs as shown in Table 4 using the same kernel. The *SVM-SubLoc* has yielded 99.7% accuracy, which is 5.3% higher than that of RBF-SVM. It has been observed that u2 LBPs, computed on a larger circle and concatenated with Haralick textures, gives more discrimination power to the classifier. MCC and F-score values of 0.98 each indicate that both the quality of prediction and accuracy of test are admirable when Halick textures are concatenated with u2 LBPs for $R=3$ and $N=24$. However, diverse results are obtained when Haralick textures are concatenated with u2 LBPs on circles of radius 1 and 2 as indicated by Q value of 0.14.

Performance of SVM-SubLoc using HarLTP

The predictions of *SVM-SubLoc* using *HarLTP* have been presented in Table 9. In individual classifiers, *poly-SVM* has yielded the highest accuracy of 94.7% using the hybrid of Haralick textures and *u2 LTPs* for $R=2$, $N=16$ and $\theta=80$. The ensemble *SVM-SubLoc* has achieved the highest accuracy of 99.4% using the hybrid of Haralick textures and *riu2 LTPs* for $R=3$, $N=24$ and $\theta=80$. The ensemble has yielded 4.7% higher accuracy than that of *poly-SVM*. *MCC* and *F*-score have yielded the highest values for the hybrid of Haralick and riu2 LTPs for $R=3$, $N=24$ and $\theta=80$. It means that prediction quality and test accuracy are promising. Maximum diversity is achieved using *HarLTP* as shown by Q value of 0.05.

3.2 Performance analysis for LOCATE datasets

In this section, Table 10 reports the performance predictions of *SVM-SubLoc* using *LTPs*, *HarLBP* and *HarLTP* features on LOCATE Endogenous and Transfected datasets. Only the best ensemble outcomes are shown here.

The complete results using these three feature sets are given in Supplementary Tables S6–S11. The *SVM-SubLoc* has achieved the

Table 9. Performance of SVM-SubLoc using HarLTP

					lin	*poly*	*RBF*	*sig*	Ensemble			
R	*N*	θ	*m*	*D*	Acc				Acc	MCC	*F*-score	*Q*-statistic
1	8	40	u2	144	92.4	92.5	92.2	85.2	98.1	0.91	0.91	0.24
1	8	40	ri	98	90.9	90.6	90.3	87.8	98.3	0.92	0.92	0.30
1	8	40	riu2	46	88.8	91.5	91.8	88.2	98.2	0.91	0.92	0.19
2	16	80	u2	512	94.3	**94.7**	94.4	92.3	99.0	0.95	0.95	0.07
2	16	80	riu2	62	93.2	93.1	93.6	90.7	99.0	0.95	0.95	0.19
3	24	80	riu2	78	93.9	93.9	93.1	90.8	**99.4**	**0.97**	**0.97**	**0.05**

Highest values obtained are represented in bold.

Table 10. Highest ensemble accuracies achieved using LOCATE Endogenous and Transfected datasets

Dataset	Feature	Ensemble accuracy	MCC	*F*-score	*Q*-statistic
Endogenous	HarLBP	**99.8**	0.98	0.99	0.10
Transfected	HarLTP	**98.7**	0.92	0.93	0.20

Highest values obtained are represented in bold.

Table 11. Performance comparison with other published work

Method	2D HeLa	LOCATE endogenous	LOCATE transfected
(Hamilton *et al.*, 2007) 5F	–	98.2 (47)	93.2 (47)
(Chebira *et al.*, 2007) 5F	95.4 (78)	–	–
(Nanni and Lumini, 2008) 5F	94.2 (107)	98.4 (107)	96.5 (81)
(Nanni *et al.*, 2010c) 10F	97.5 (322)	–	–
(Nanni *et al.*, 2010a) 5F	95.8 (305)	99.5 (305)	97.0 (305)
SVM-SubLoc using *HarLBP*	**99.7** (581)	**99.8** (36)	98.5 (44)
SVM-SubLoc using *HarLTP*	**99.4** (78)	99.6 (62)	**98.7** (78)
SVM-SubLoc using *LTPs*	**99.0** (52)	95.6 (36)	93.6 (36)

5F and 10F represent 5-fold and 10-fold, respectively. Highest values obtained are represented in bold.

highest accuracy of 99.8% using *HarLBP* features for LOCATE Endogenous dataset. However, *SVM-SubLoc* has yielded 98.7% accuracy using *HarLTP* for LOCATE Transfected dataset.

3.3 Comparison with existing approaches

In Table 11, we have carried out a performance comparative analysis of the proposed *SVM-SubLoc* approach with previously well-known approaches for *2D HeLa* and the two LOCATE datasets.

The accuracy of 95.4% is obtained by the proposed approach for *2D HeLa* dataset in (Chebira *et al.*, 2007). Nanni and Lumini have reported accuracies of 94.2, 98.4 and 96.5% for the *2D HeLa*, LOCATE Endogenous and Transfected datasets, respectively (Nanni and Lumini, 2008). In an another paper, Nanni *et al.* (2010c) have reported the highest accuracy 97.5% for *2D HeLa* dataset. Nanni *et al.* (2010a) have also reported 95.8% accuracy for 2*D HeLa* dataset, 99.5% for LOCATE Endogenous and 97.0% accuracy for LOCATE Transfected dataset. On the other hand, our proposed *SVM-SubLoc* approach has yielded 99.7% accuracy using *HarLBP* for *2D HeLa* dataset that is 2.20% higher than the highest accuracy reported in Nanni *et al.* (2010c).

In addition, using *HarLBP* our approach has yielded the accuracy of 99.8% for LOCATE Endogenous dataset, which is 0.3% higher than the highest accuracy obtained by Nanni *et al.* (2010a). Similarly, using Har*LTP*, our approach achieved 98.7% accuracy for LOCATE Transfected dataset that is 1.7% higher than that of the proposed technique by Nanni *et al.* (2010a).

The performance of *SVM-SubLoc* approach is enhanced due to the two level ensembles; one is at the features level and the other is at the classifiers decision level. At features level, we have constructed the hybrid features by concatenating different individual feature sets. These features have improved the discrimination power of the features. At the classifiers decision level, we have combined the predictions of the utilized *SVMs* through the majority-voting scheme.

4 CONCLUSIONS

We have presented a simple, accurate and effective prediction model for protein subcellular location images from *2D HeLa* and the two LOCATE datasets. The proposed *SVM-SubLoc* approach is reliable and computationally efficient. We have utilized both spatial and transform domain features. It has been shown that the performance of *SVM-SubLoc* is better compared with the individual classifiers using the hybrid features particularly *HarLBP* and *HarLTP* in spatial domain. The prediction accuracy has reached to 99.7% using *HarLBP* with features of 581 dimensions. On the other hand, the accuracy is 99.4% using *HarLTP* but the dimensionality of the feature space is reduced to 78 only, which is an effective reduction. These features are computationally more reasonable in terms of cost along with better discrimination capability compared with other texture based feature extraction strategies. In individual features, *LTPs* outperforms both hybrid and other individual features in terms of dimensionality of the feature space that is only 52. Even though the accuracy is 99.0%, which is a little less than that of *HarLBP*, it is preferable to use *LTPs* because of less computational cost. Additionally, these three feature sets have also performed well for the two LOCATE datasets. The hybrid features have brought a significant improvement in performance. This is due to the fact that the discrimination powers of both the feature spaces are utilized by *SVM-SubLoc*.

The comparative analysis highlights the improved performance of our proposed *SVM-SubLoc* approach in terms of both increased accuracy and reduced dimensionality of the feature space over existing well-known approaches.

ACKNOWLEDGEMENTS

We are thankful to Dr. Loris Nanni Associate Researcher, Department of Information Engineering - University of Padua, Italy for his help in implementation.

Funding: Higher Education Commission of Pakistan under the indigenous PhD scholarship program 17-5-4(Ps4-124)/HEC/Sch/2008/.

Conflict of Interest: none declared.

REFERENCES

Boland,M.V. *et al.* (1998) Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, **33**, 366–375.

Boland,M.V. and 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.

Chebira,A. *et al.* (2007) A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, **8**, 210.

Gunn,S.R. (1998) Support Vector Machines for Classification and Regression. *Technical Report*. Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton, Southampton.

Hamilton,N.A. *et al.* (2006) Automated Sub-Cellular Phenotype Classification: An Introduction and Recent Results. *The 2006 Workshop on Intelligent Systems for Bioinformatics* (*WISB 2006)*. Australian Computer Society, Inc., Hobart, Australia, pp. 67–72.

Hamilton,N.A. *et al.* (2007) Fast automated cell phenotype image classification. *BMC Bioinformatics*, **8**, 110.

Haralick,R.M. (1979) Statistical and Structural Approaches to Texture. In *Proceedings of the IEEE*, Vol. 67. IEEE, pp. 786–804.

Hayat,M. and Khan,A. (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. *J. Theor. Biol.*, **271**, 10–17.

Hu,Y. and Murphy,R.F. (2004) Automated interpretation of subcellular patterns from immunofluorescence microscopy. *J. Immunol. Methods*, **290**, 93–105.

Khan,A. *et al.* (2008) Machine learning based adaptive watermark decoding in view of an anticipated attack. *Patt. Recognit.*, **41**, 2594–2610.

Khan,A. *et al.* (2011) CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of Pseudo amino acid composition. *Comput. Biol. Chem.*, **35**, 218–229.

Lin,C.-C. *et al.* (2007) Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization. *Bioinformatics*, **23**, 3374–3381.

Majid,A. *et al.* (2006) Combining support vector machines using genetic programming. *Int. J. Hybrid Intell. Syst.,* **3**, 109–125.

Murphy,R.F. *et al.* (2000) 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, pp. 251–259.

Murphy,R.F. *et al.* (2002) Robust classification of subcellular location patterns in fluorescence microscope images. In *Proceedings of the 2002 12th IEEE International Workshop on Neural Networks for Signal Processing*. IEEE, pp. 67–76.

Murphy,R.F. *et al.* (2003) Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *J. VLSI Signal Process.*, **35**, 311–321.

Nanni,L. *et al.* (2009) Automated cell phenotype image classification combining different methods. In *Proceeding of the Workshop on Automated Interpretation and Modeling of Cell Images (ICML-UAI-COLT 2009)*. McGill University, Montreal QC, Canada.

Nanni,L. and Lumini,A. (2008) A reliable method for cell phenotype image classification. *Art. Intell. Med.*, **43**, 87–97.

Nanni,L. *et al.* (2010a) Novel features for automated cell phenotype image classification. *Adv. Comput. Biol. Adv. Exp. Med. Biol.*, **680**, 207–213.

Nanni,L. *et al.* (2010b) 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, pp. 369–375.

Nanni,L. *et al.* (2010c) Fusion of systems for automated cell phenotype image classification. *Exp. Syst. Appl.*, **37**, 1556–1562.

Ojala,T. *et al.* (1996) A comparative study of texture measures with classification based on feature distribution. *Patt. Recognit.*, **29**, 51–59.

Srinivasa,G. *et al.* (2006) Adaptive multiresolution techniques for subcellular protein location classification. *IEEE Int. Conf. Acoustics Speech Signal Process*, **5**, 14–19.

Tan,X. and Triggs,B. (2007) Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *Proceedings of the 3rd International Conference on Analysis and Modeling of Faces and Gestures*. Springer, Rio de Janeiro, Brazil, pp. 168–182.