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

asif@pieas.edu.pk*

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 paper, the predictions of various 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 to 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 to existing approaches.

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

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 in 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; Murphy, et al., 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 2D 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. have trained BPNN using enhanced feature sets consisting of Haralick textures, Zernike moments, and morphological features (Murphy, et al., 2003). Hamilton et al. have employed SVM using various feature extraction strategies including Threshold Adjacency Statistics (TASs), Zernike moments as well as a hybrid of TASs and Haralick textures (Hamilton, et al., 2007). Chebira et al. have developed automated classification system for protein subcellular location images in multi-resolution subspaces (Chebira, et al., 2007). They have employed ANN at different decomposition levels to obtain the classification results, which are then combined through weight assignment. Nanni et al. have utilized random subspace of Levenberg-Marquardt neural networks and AdaBoost learning algorithm (Nanni, et al., 2009). 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. 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 (LBP), Local Ternary Patterns (LTP), and Threshold Adjacency Statistics for 2D HeLa and LOCATE mouse protein datasets (Nanni, et al., 2010a).

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 to the existing approaches. We thus develop both individual and hybrid features based classification approaches for the prediction of protein subcellular localization. In the proposed approach, we employ Haralick textures, Zernike moments, LBP, LTP, and TAS 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 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

*To whom correspondence should be addressed.
the image up to 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 paper is organized as follows: Dataset 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 Dataset

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 x 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 x 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 12 and 13, respectively.

2.2 Feature Extraction Strategies

In this work, we have employed various texture based feature extraction strategies such as Haralick textures, Zernike moments, LBP, 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 \( \theta \) (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 thirteen 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 to other texture based features (Hu and Murphy, 2004). Zernike moments of an image can be calculated using Equation 1 (Boland, et al., 1998).

\[
Z_{nl}(x, y) = \frac{n + 1}{\pi} \sum_{k, m} V_{nl}(x, y) I(x, y)
\]

(1)

where \( I(x, y) \) represents the pixel intensity at position \((x, y)\), \( x^2 + y^2 \leq 1 \), \( 0 \leq l \leq n \), \( n - l \) is even. \( V_{nl}(x, y) \) indicates complex conjugate of the Zernike Polynomial of degree \( n \) and angular dependence \( \theta \). Its value can be computed as follows:

\[
V_{nl}(x, y) = \sum_{m=0}^{(n-l)/2} (-1)^m \frac{(x^2 + y^2)^{n/2 - m}}{m! \left(\frac{n}{2} - m\right)! \left(\frac{n}{2} + m\right)!} e^{i\theta (n - m)}
\]

(2)

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

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 sub-images. 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 (LBP)

Local Binary Patterns is a texture based feature extraction strategy for gray level patterns in an image, proposed by (Ojala, et al., 1996). LBP 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; Nanni, et al., 2010b; Nanni and Lumini, 2008). The procedure of obtaining LBP code is depicted in Figure 1.
Protein Subcellular Localization of Fluorescence Imagery using Spatial and Transform domain Features

Figure 1 Procedure of LBPs code generation

\[ s(x) = \begin{cases} 
1 & \text{if } x \geq \theta \\
0 & \text{otherwise}
\end{cases} \]

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)

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 \( \theta \) as given by Equation 4.

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

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; Nanni, et al., 2010b).

\[ LBPR,R = \sum_{p=0}^{R-1} s(p, R) 2^p \]

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)

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.

\[ \text{Threshold Adjacency Statistics: 0 to 8 neighboring white pixels of a central white pixel in a 3 x 3 neighborhood} \]

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 \( \mu \) to 255, \( \mu - \theta \) to 255, and \( \mu + \theta \) to 255 where \( \mu \) is the average pixel intensity and \( \theta \) is the user defined threshold. The value of threshold \( \theta \)
should not be less than 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 Support Vector Machine (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, 2010; Majid, et al., 2006). It was developed for binary classification problems. However, to employ SVM for multi-classification problems, a straightforward approach is to reduce the multi-classification to a series of binary classifications through one-versus-rest mechanism. For a \( k \)-class classification problem, \( k \) SVMs are built where the \( i^{th} \) SVM is trained on every instance in the \( i^{th} \) 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 Threshold Adjacency Statistics 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.

![Figure 4](image)

**Figure 4** Framework of the proposed SVM-SubLoc

### 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.

\[
Q_{avg} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=1}^{L} Q_{i,k}
\]

(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.

a. 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.

<table>
<thead>
<tr>
<th>L</th>
<th>D</th>
<th>Acc</th>
<th>lin</th>
<th>poly</th>
<th>RBF</th>
<th>sig</th>
<th>MCC</th>
<th>F-Score</th>
<th>Q-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>26</td>
<td>75.6</td>
<td>75.9</td>
<td>76.4</td>
<td>69.6</td>
<td>87.7</td>
<td>0.59</td>
<td>0.60</td>
<td>0.29</td>
</tr>
<tr>
<td>1</td>
<td>104</td>
<td>80.5</td>
<td>80.9</td>
<td>81.7</td>
<td>80.3</td>
<td>92.5</td>
<td>0.71</td>
<td>0.72</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>416</td>
<td>81.9</td>
<td>83.2</td>
<td>84.1</td>
<td>79.8</td>
<td><strong>93.2</strong></td>
<td>0.73</td>
<td><strong>0.74</strong></td>
<td>0.20</td>
</tr>
<tr>
<td>3</td>
<td>1664</td>
<td>77.3</td>
<td>78.3</td>
<td>80.6</td>
<td>77.9</td>
<td>90.2</td>
<td>0.65</td>
<td>0.66</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>6656</td>
<td>79.5</td>
<td>80.6</td>
<td>80.7</td>
<td>79.5</td>
<td>92.6</td>
<td>0.72</td>
<td>0.73</td>
<td>0.23</td>
</tr>
</tbody>
</table>

The individual RBF-SVM has achieved the highest success rates at all decomposition levels. Particularly, its highest accuracy 84.1% is observed at 2nd 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 2nd 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 2nd 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 1st decomposition level with the Q value 0.18.

b. 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.

<table>
<thead>
<tr>
<th>L</th>
<th>D</th>
<th>Acc</th>
<th>lin</th>
<th>poly</th>
<th>RBF</th>
<th>sig</th>
<th>MCC</th>
<th>F-Score</th>
<th>Q-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>49</td>
<td>28.3</td>
<td>43.8</td>
<td>46.5</td>
<td>15.5</td>
<td>46.7</td>
<td>-0.02</td>
<td>0.16</td>
<td>0.02</td>
</tr>
<tr>
<td>1</td>
<td>196</td>
<td>35.9</td>
<td>50.3</td>
<td>52.2</td>
<td>24.4</td>
<td>57.8</td>
<td>0.12</td>
<td>0.23</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>784</td>
<td>39.0</td>
<td>53.4</td>
<td>56.2</td>
<td>15.1</td>
<td><strong>58.5</strong></td>
<td>0.15</td>
<td><strong>0.25</strong></td>
<td><strong>0.01</strong></td>
</tr>
<tr>
<td>3</td>
<td>3156</td>
<td>50.2</td>
<td>52.0</td>
<td>60.5</td>
<td>11.3</td>
<td>54.5</td>
<td>0.10</td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>4</td>
<td>12544</td>
<td>44.0</td>
<td>48.0</td>
<td><strong>67.8</strong></td>
<td>11.0</td>
<td>56.7</td>
<td>0.14</td>
<td>0.24</td>
<td>0.11</td>
</tr>
</tbody>
</table>

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 4th level. Though, SVM-SubLoc has yielded the best accuracy 58.5% at 2nd level due to the maximum diversity at this level as indicated by the Q value 0.01. However, prediction quality and accuracy of the test are not reasonably good as shown by the MCC and F-Score values at 2nd decomposition level.

c. Performance of SVM-SubLoc using TASs

Table 3 demonstrates the predictions of SVM-SubLoc using TASs for gray images without DWT.

<table>
<thead>
<tr>
<th>L</th>
<th>D</th>
<th>Acc</th>
<th>lin</th>
<th>poly</th>
<th>RBF</th>
<th>sig</th>
<th>MCC</th>
<th>F-Score</th>
<th>Q-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>27</td>
<td>77.6</td>
<td>80.3</td>
<td><strong>81.0</strong></td>
<td>70.7</td>
<td><strong>91.6</strong></td>
<td>0.69</td>
<td><strong>0.70</strong></td>
<td><strong>0.31</strong></td>
</tr>
</tbody>
</table>

In individual base classifiers, RBF-SVM has achieved the highest accuracy 81.0% compared to other SVMs. The success rate of SVM-SubLoc is 91.6%, which is 10.5% higher than that of RBF-SVM. The increased ensemble accuracy shows the significance of the ensemble classifier. MCC value 0.69 shows good quality of the prediction whereas F-Score value 0.70 indicates fine accuracy of the performed test. The Q value 0.31 reveals that results have 69% diversity. It has been analyzed during the experiments that TASs have produced the most significant results for θ = 140. The discriminating capability of these features at this threshold is enhanced. The results achieved at other threshold values are presented in supplementary Table 1 for comparison.
d. 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 88.0% using \( u_2 \) LBPs for \( R=3 \) and \( N=24 \). However, SVM-SubLoc has yielded the highest accuracy 95.5% using \( u_2 \) 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 0.25. The best values of MCC (0.81) and F-Score (0.82) are also obtained using \( u_2 \) LBPs.

e. 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 \( \theta \) 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 \( riu_2 \) LTPs. The highest accuracy 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 compared to other feature sets. As is evident from the \( Q \) value 0.15, we observed the maximum diversity using \( riu_2 \) LTPs for \( R=3 \) and \( N=24 \) and that is the main reason why the ensemble accuracy is high. An MCC value 0.95 and F-Score value 0.95 also indicate that discrimination power of \( riu_2 \) 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 \( \theta \) performs well. However, at larger circles, the value of \( \theta \) 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.

a. 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 2nd decomposition level has been found to be the best level for discriminating subcellular location images. However, most diverse results are obtained at 1st level as indicated by $Q$ value 0.12. Highest values of MCC 0.73 and F-Score 0.74 are also achieved at 2nd 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 to using their individual constituents. However, SVM-SubLoc has yielded the same accuracy as yielded by the ensemble using Haralick textures.

b. 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 87.9%, which is further enhanced by the ensemble SVM-SubLoc up to 96.2%.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$D$</th>
<th>Acc</th>
<th>MCC</th>
<th>F-Score</th>
<th>$Q$-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>53</td>
<td>87.0</td>
<td>87.9</td>
<td>86.0</td>
<td>83.0</td>
</tr>
</tbody>
</table>

This shows 8.3% improvement in the accuracy. The $Q$ value 0.32 indicates sufficient diversity among classifiers. Quality of the prediction and accuracy of the test are quite good as revealed by MCC value 0.83 and F-Score value 0.84, respectively. The results at other threshold values are presented in supplementary Table 2.

c. 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 94.4% for $\theta=3$ and $N=24$. The obtained accuracy is 17.9% higher than the highest accuracy using Haralick textures (at 0th level) as given in Table 1 and 63% higher than the highest accuracy using LBP's 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 $u2$ LBP's, computed on a larger circle and concatenated with Haralick textures, gives more discrimination power to the classifier. MCC and F-Score values 0.98 each indicate that both the quality of prediction and accuracy of test are admirable when Haralick textures are concatenated with $u2$ LBP's for $\theta=3$ and $N=24$. However, diverse results are obtained when Haralick textures are concatenated with $u2$ LBP's on circles of radius 1 and 2 as indicated by $Q$ value 0.14.

<table>
<thead>
<tr>
<th>$R$</th>
<th>$N$</th>
<th>$\theta$</th>
<th>$m$</th>
<th>$D$</th>
<th>Acc</th>
<th>MCC</th>
<th>F-Score</th>
<th>$Q$-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>2</td>
<td>85</td>
<td>92.1</td>
<td>92.4</td>
<td>92.5</td>
<td>88.5</td>
<td>99.3</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>4</td>
<td>82</td>
<td>89.2</td>
<td>89.9</td>
<td>89.0</td>
<td>84.9</td>
<td>97.0</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>36</td>
<td>88.3</td>
<td>90.2</td>
<td>89.6</td>
<td>87.9</td>
<td>97.5</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>2</td>
<td>269</td>
<td>93.2</td>
<td>93.9</td>
<td>94.4</td>
<td>91.7</td>
<td>99.0</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>4</td>
<td>4142</td>
<td>81.6</td>
<td>81.6</td>
<td>78.5</td>
<td>81.0</td>
<td>92.5</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>8</td>
<td>44</td>
<td>92.5</td>
<td>93.2</td>
<td>93.7</td>
<td>90.8</td>
<td>98.8</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>2</td>
<td>581</td>
<td>93.2</td>
<td>93.6</td>
<td>94.4</td>
<td>90.9</td>
<td>97.7</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>4</td>
<td>52</td>
<td>91.8</td>
<td>92.6</td>
<td>92.8</td>
<td>89.7</td>
<td>99.3</td>
</tr>
</tbody>
</table>

Table 9 Performance of SVM-SubLoc using HarLTP

<table>
<thead>
<tr>
<th>$R$</th>
<th>$N$</th>
<th>$\theta$</th>
<th>$m$</th>
<th>$D$</th>
<th>Acc</th>
<th>MCC</th>
<th>F-Score</th>
<th>$Q$-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>40</td>
<td>144</td>
<td>92.4</td>
<td>92.5</td>
<td>92.2</td>
<td>85.2</td>
<td>98.1</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>40</td>
<td>98</td>
<td>90.9</td>
<td>90.6</td>
<td>90.3</td>
<td>87.8</td>
<td>98.3</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>40</td>
<td>146</td>
<td>88.8</td>
<td>91.5</td>
<td>91.8</td>
<td>88.2</td>
<td>98.2</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>80</td>
<td>512</td>
<td>94.3</td>
<td>94.7</td>
<td>94.4</td>
<td>92.3</td>
<td>99.0</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>80</td>
<td>62</td>
<td>93.2</td>
<td>93.1</td>
<td>93.6</td>
<td>90.7</td>
<td>99.0</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>80</td>
<td>78</td>
<td>93.9</td>
<td>93.9</td>
<td>93.1</td>
<td>90.8</td>
<td>99.4</td>
</tr>
</tbody>
</table>

d. 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 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 99.4% using the hybrid of Haralick textures and $rin2$ LTPs for $R=3$, $N=24$, and $\theta=80$. The ensemble has yielded 4.6% higher accuracy than that of poly-SVM. MCC and F-Score have yielded the highest values for the hybrid of Haralick and $rin2$ LTPs for
HarLBP achieved accuracies of 94.2%, 98.4%, and 96.5% for the Transfected dataset (Nanni, et al., 2010a). On the other hand, our proposed approach has also reported 95.8% accuracy for LOCATE Endogenous dataset. However, SVM-SubLoc has yielded 98.7% accuracy using HarLTP for LOCATE Transfected dataset.

### 3.2 Performance Analysis for LOCATE dataset

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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Feature</th>
<th>Ensemble</th>
<th>MCC</th>
<th>F-Score</th>
<th>Q-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endogenous</td>
<td>HarLBP</td>
<td>99.8</td>
<td>0.98</td>
<td>0.99</td>
<td>0.10</td>
</tr>
<tr>
<td>Transfected</td>
<td>HarLTP</td>
<td>98.7</td>
<td>0.92</td>
<td>0.93</td>
<td>0.20</td>
</tr>
</tbody>
</table>

The complete results using these three feature sets are given in supplementary Tables 6, 7, 8, 9, 10, and 11. The SVM-SubLoc has achieved the 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 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. have reported the highest accuracy 97.5% for 2D HeLa dataset (Nanni, et al., 2010c). Nanni et al. have also reported 95.8% accuracy for 2D HeLa dataset, 99.5% for LOCATE Endogenous, and 97.0% accuracy for LOCATE Transfected dataset (Nanni, et al., 2010a). 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>2D HeLa</th>
<th>LOCATE Endogenous</th>
<th>LOCATE Transfected</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Hamilton, et al., 2007) 5F</td>
<td>-</td>
<td>93.2 (47)</td>
<td>92.2 (47)</td>
</tr>
<tr>
<td>(Chebira, et al., 2007) 5F</td>
<td>95.4 (78)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Nanni and Lumini, 2008) 5F</td>
<td>94.2 (107)</td>
<td>98.4 (107)</td>
<td>96.5 (81)</td>
</tr>
<tr>
<td>(Nanni, et al., 2010c) 10F</td>
<td>97.5 (322)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Nanni, et al., 2010a) 5F</td>
<td>95.8 (305)</td>
<td>99.5 (305)</td>
<td>97.0 (305)</td>
</tr>
<tr>
<td>SVM-SubLoc using HarLBP</td>
<td>99.7 (581)</td>
<td>99.8 (36)</td>
<td>98.5 (44)</td>
</tr>
<tr>
<td>SVM-SubLoc using HarLTP</td>
<td>99.4 (78)</td>
<td>99.6 (62)</td>
<td>98.7 (78)</td>
</tr>
<tr>
<td>SVM-SubLoc using LTPs</td>
<td>99.0 (52)</td>
<td>95.6 (36)</td>
<td>93.6 (36)</td>
</tr>
</tbody>
</table>

5F and 10F represent 5-fold and 10-fold respectively.

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 HarLTP, 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.

### 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 to 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 to 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

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

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES


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.


