



# Pattern analysis of protein images from fluorescence microscopy using Gray Level Co-occurrence Matrix



Muhammad Tahir

College of Computing and Informatics, Saudi Electronic University, Al-Madinah Branch, Saudi Arabia

Received 18 July 2016; accepted 20 December 2016

Available online 29 December 2016

## KEYWORDS

Texture analysis;  
Protein subcellular localization;  
GLCM;  
Haralick textures;  
Support vector machine;  
Correlation based feature selection

**Abstract** Extraction of useful and discriminative information from fluorescence microscopy protein images is a challenging task in the field of machine learning and pattern recognition.

Gray Level Co-occurrence Matrix (GLCM) was among the first methods developed for textural analysis, which holds information of intensity distribution as well as the respective distance of intensity levels in the original image. In this paper, several GLCMs are constructed with different quantization levels for different values of offset  $d$ . Haralick descriptors are extracted from each GLCM, which are then utilized to train support vector machines. The final output is obtained through the majority voting scheme. Hybrid models from different individual feature spaces have also been constructed. Additionally, Correlation-based Feature Selection (CFS) is performed to extract the most useful features from the hybrid models.

The empirical analysis reveals that varying the value of parameter  $d$  causes the GLCM to extract different information from a particular fluorescence microscopy image. Hence, producing diversified co-occurrence matrices for same images. Similarly, using more quantization levels for constructing a GLCM generates informative and discriminative features for the classification phase. Furthermore, CFS has significantly reduced the feature space dimensionality achieving almost the same accuracy as full feature space.

The performance of the proposed system is validated using three benchmark datasets including HeLa (99.6%), CHO (100%), and LOCATE Endogenous (100%) datasets. It is anticipated that GLCM is still an efficient technique for pattern analysis in the field of bioinformatics and computational biology as well as might be helpful in drug discovery related applications.

© 2016 The Author. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

E-mail address: [m.tahir@seu.edu.sa](mailto:m.tahir@seu.edu.sa)

Peer review under responsibility of King Saud University.



## 1. Introduction

Cell plays critical role in the existence of life. Proteins are the basic building blocks of all cells and are responsible to execute almost all cellular functions (Alberts et al., 2002; Cooper,

2000). The understanding of protein functions is among the top priorities in the field of biological sciences. Subcellular localization of proteins in a cell is one of the key characteristics of proteins. Exact localization of a particular protein delivers precious information about the functionality of a protein under various circumstances (Boland et al., 1998; Chebira et al., 2007). Predicting the precise location of a protein, for example, may convey useful information during the drug discovery process. Further, the efficiency of a particular drug, in treating a particular disease, may be estimated by accurate prediction of subcellular localization. Similarly, doctors may detect a disease at its earlier stages upon having the knowledge of accurate localization of a particular protein. The current trends, in advancement of imaging technologies, are producing a huge amount of cell images for drug discovery (Chen et al., 2006; Newberg et al., 2009). However, analysis of these images for classification in traditional ways is laborious, error prone and almost impossible. Therefore, automated systems are needed to classify these images accurately, timely and reliably. During the past decade, researchers have been developing Bioinformatics based automated systems for the classification of protein subcellular localization images (Chen et al., 2006; Chen and Li, 2013; Glory and Murphy, 2007; Li et al., 2012; Nanni et al., 2013). These systems are capable of recognizing and classifying major protein compartments by utilizing numerical descriptions of fluorescence microscopy images through various machine learning techniques (Boland and Murphy, 2001; Chebira et al., 2007; Nanni et al., 2010a; Tahir et al., 2012).

In this regards, Nanni et al. have introduced a random subspace based model for selecting local binary and ternary patterns having high variance. A support vector machine is trained to classify protein images into various classes (Nanni et al., 2010b). Srinivasa et al. have extracted Haralick textures and morphological features from images at the sub-bands and then K-means algorithm is applied to these features separately at sub-bands to classify protein images. Final prediction is obtained by combining these individual predictions through weighting strategy (Srinivasa et al., 2006). Similarly, Nanni et al. have proposed a model based on the ensemble of Levenberg–Marquardt neural network and AdaBoost algorithm using random subspace of numerous hybrid feature sets. The decisions of the two ensembles are fused through sum rule (Nanni et al., 2010c). Likewise, Murphy et al. have proposed a back propagation neural network based model, which classifies protein subcellular localization images into various classes using Haralick textures, Zernike moments, and morphological features (Murphy et al., 2003). Tscherepanow et al. have utilized various feature extraction strategies in conjunction with a modified version of fuzzy ARTMAP (SFAM) as a classification algorithm. The feature extraction techniques include pattern spectra, fractal features, histogram based features, Zernike moments, and region dependent texture features (Tscherepanow et al., 2008). Tahir et al. have developed *SVM-SubLoc* (Tahir et al., 2012), which constructed the feature spaces in different sub-bands of the original image and then various support vector machines are trained using the extracted features. The decisions of the individual support vector machines are combined through the majority voting technique. In another work, Tahir et al. have proposed *RF-SubLoc* (Tahir et al., 2013) prediction system that utilized the Synthetic Minority Oversampling Technique to oversam-

ple the original samples and then utilized the oversampled samples to train Random Forest classifier for the classification of protein images.

Literature reveals that GLCM is the focus of many researchers in the field of computer vision, pattern recognition and machine learning (Chen et al., 2009; Gelzinis et al., 2007; Mitrea et al., 2012; Walker et al., 2003). It is among the most primitive techniques for texture based analysis. These researchers have adopted different approaches for extracting information from GLCM matrices.

In this paper, GLCM based Haralick textural features are utilized to address the challenging problem of protein subcellular localization. The focus of our research is to analyze GLCM for its discriminative capability regarding different values of offset parameter  $d$  against a particular quantization level. Results showed the effectiveness of GLCM based texture analysis for protein subcellular localization images. We also utilized CFS in order to extract the most informative features from the full feature space. Empirical analysis reveals that feature selection efficiently reduced the feature space while keeping the discriminative power intact.

Rest of the paper is organized as follows. Section 2 presents materials and methods. Section 3 analyzes the simulation results. Section 4 presents the comparative analysis. Section 5 draws the conclusive remarks.

## 2. Materials and methods

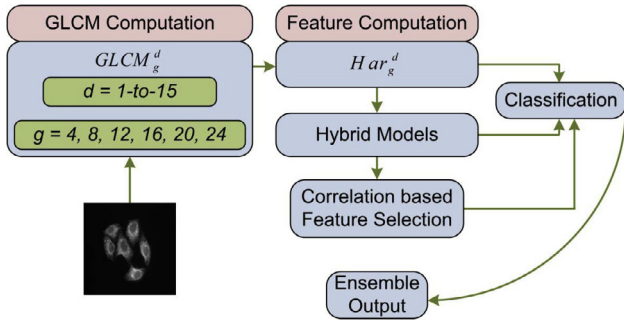
In this section, we first discuss the datasets that are used to assess the performance of the proposed algorithm. Next, the proposed prediction system is discussed. Then, the feature extraction technique is elaborated in detail. The hybrid models and feature selection are also discussed toward the end of this section.

### 2.1. Datasets

Three benchmark protein image datasets have been utilized to assess the performance of our proposed method. These include 2D HeLa (Boland and Murphy, 2001), CHO (Lin et al., 2007), and LOCATE Endogenous (Nanni et al., 2010a) datasets. The cells in HeLa dataset, comprised of 862 images, are distributed in 10 different classes including Actin Filaments, Endosome, Endoplasmic Reticulum, Golgi Giantin, Golgi GPP130, Lysosome, Microtubules, Mitochondria, Nucleolus, and Nucleus. CHO dataset seizes 668 protein images categorized in 8 unique classes, which include Actin, Endoplasmic Reticulum, Golgi, Microtubule, Mitochondria, Nucleolus, Nucleus, and Peroxisome. The third dataset, LOCATE Endogenous contains 502 protein images and comprised of 10 classes, which are Actin, Endosome, Endoplasmic Reticulum, Golgi, Lysosome, Microtubule, Mitochondria, Nucleus, Peroxisome, and Plasma Membrane.

### 2.2. The proposed prediction system

Fig. 1 demonstrates the proposed *GLCM-SubLoc* prediction system. The input image is first quantized to the required number of gray levels. Then for a particular value of offset parameter  $d$ , a GLCM is constructed along four directions: horizontal, vertical, diagonal, and off-diagonal as shown in

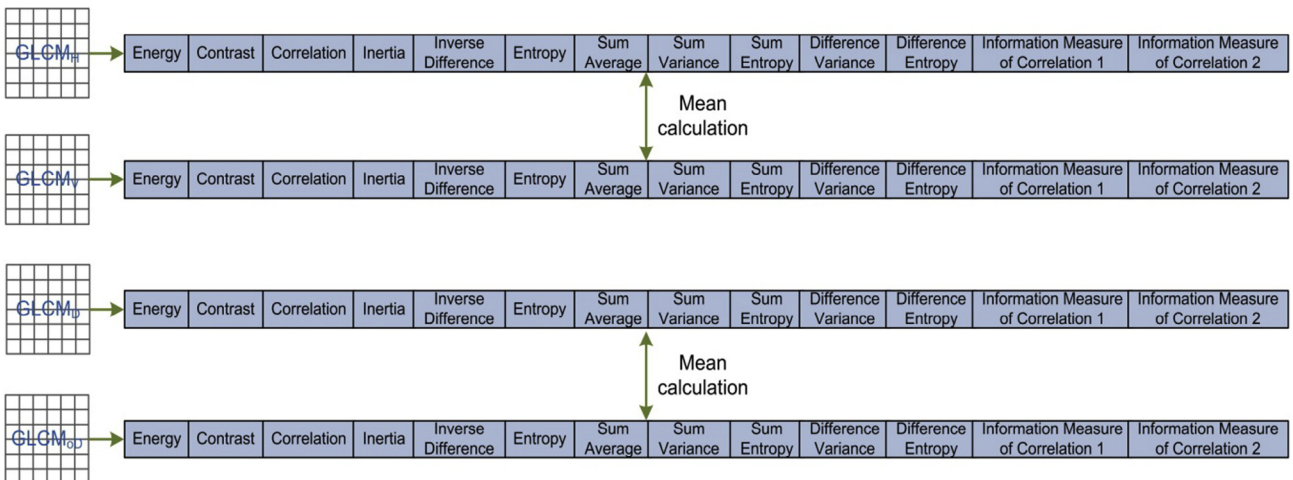


**Figure 1** The proposed *GLCM-SubLoc* prediction system.

**Fig. 2.** Next, features are extracted from each of the four GLCMs separately. Then features from horizontal GLCM are averaged with the features from vertical directional GLCM. Similarly, the features from diagonal GLCM are averaged with the features from off-diagonal GLCM. The final feature space is constructed by concatenating the two averaged feature spaces.

From this point onward, a GLCM constructed with a particular quantization level  $g$  and offset value  $d$  is referred to as  $GLCM_g^d$  where different values of  $g$  considered in this paper are 4, 8, 12, 16, 20, and 24 whereas values of  $d$  are from the set 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15.  $Har_g^d$  indicates Haralick features extracted from a  $GLCM_g^d$ .

Hence, against a single gray level value there are fifteen values of the offset parameter  $d$  producing fifteen different feature spaces, which are then utilized to train fifteen support vector machines. These results are combined through the majority voting scheme to produce the final output. Besides, various hybrid models are also constructed from different individual feature spaces. In order to select informative features from these hybrid models, we incorporated Correlation-based Feature Selection (CFS) technique. Moreover, support vector machines are trained using the hybrid as well as CFS based selected feature spaces for comparison purpose.



**Figure 2** Feature vectors for GLCM along four different directions including: horizontal, vertical, diagonal, and off-diagonal.

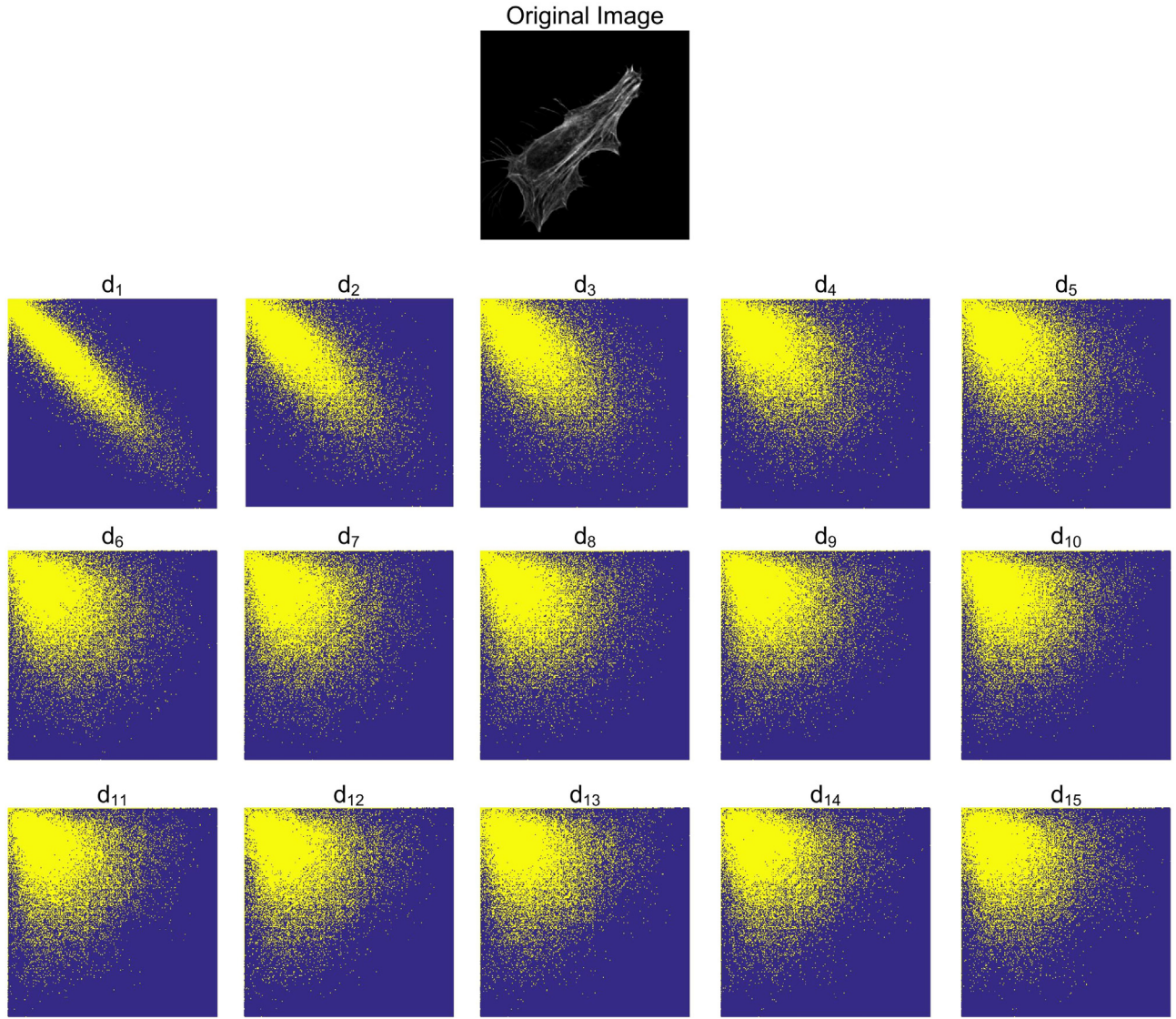
### 2.3. Feature extraction strategy

Haralick texture features, extracted from various GLCMs are utilized in the development of *GLCM-SubLoc* prediction system. GLCM maintains co-occurring frequency information of two pixels in a particular relation. This relationship is defined by two parameters: the offset  $d$  and orientation  $\theta$ . In the current study, we used  $d = 1, 2, 3, \dots, 15$  and  $\theta = 0^\circ, 45^\circ, 90^\circ, \text{ and } 135^\circ$ . GLCM matrices obtained with different  $(d, \theta)$  combinations capture different information from the images. This fact is evident from Fig. 3 where GLCMs constructed with different value of  $d$  portrays different information. These GLCMs are constructed for the Actin protein image from HeLa dataset, shown in Fig. 3.

Consequently, the extracted features from these GLCM matrices possess diversified information related to textural appearance of images. The size of GLCM is dependent upon the gray level values held by an image and not on the image size itself. Let  $I$  be an image with  $N$  gray levels, the GLCM for image  $I$  will be an  $N$ -by- $N$  dimensional matrix. This GLCM, at location  $(i, j)$ , records the number of times two intensity levels  $i$  and  $j$  co-occur in the image  $I$  at distance  $d$  from each other at orientation  $\theta$ . GLCM, of an image  $I$  with  $r$  rows,  $c$  columns and offset  $(d_x, d_y)$ , can be represented mathematically as given in Eq. (1).

$$GLCM_{d_x, d_y}(i, j) = \sum_{p=1}^r \sum_{q=1}^c \begin{cases} 1, & \text{if } I(p, q) = i \text{ and } I(p + d_x, q + d_y) = j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

However, the GLCM is required to hold the probability rather than the count of the co-occurrence of any two intensities. Therefore, the GLCM entries are transformed so that they indicate probabilities. For this purpose, the number of times a particular combination of intensities occurs is divided by the total number of possible outcomes, in order to obtain probabilities. These probabilities are not true probabilities rather these are approximations. True probabilities require continuous values whereas the entries of GLCM are discrete values. Eq. (2) is used to transform a GLCM into approximate probabilities.



**Figure 3** GLCMs are constructed with varying values of parameter  $d$  for protein “original image” at the top.

$$P(i, j) = \frac{V_{ij}}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} V_{ij}} \quad (2)$$

Here  $i$  and  $j$  represent the row and column of a GLCM,  $V(i, j)$  indicates the total number of co-occurrences of intensity levels  $i$  and  $j$  whereas the summation in the denominator is the total number of co-occurrences of  $(i, j)$  (Do-Hong et al., 2010; Haralick, 1979).

One important issue related to GLCM is its size, which depends upon the gray levels of an image. For different combinations of  $(d, \theta)$ , there are four different possible GLCM matrices. Images, with a large number of gray levels, produce GLCM matrices that have huge amount of data storage requirements. Therefore, images are usually quantized to a low number of gray levels. However, reducing gray levels in an image may result in loss of texture information. Further, with a low number of gray levels homogeneity is more likely in such images (Chaddad et al., 2011; Liang and Malm,

2012). On the other hand, a large number of gray levels possess more detailed information for discrimination among patterns. However, this results in GLCM matrices, which are more sparse in nature and likely to have large vacancy level.

In this connection, we explore the effect of various quantization levels in conjunction with different values of offset parameter  $d$  on the information capturing capability of GLCM for fluorescence microscopy protein images. Distinct quantization levels as well as different values of offset parameter  $d$  result in different GLCM matrices and therefore, their produced feature spaces are diverse in nature. The occupancy level of GLCM is also affected by smooth image regions. For homogeneous images, the resultant GLCM will have higher count for some indices and most of the entries will correspond to zero. In such cases, the constructed GLCM should be based on distant neighbors (Srinivasan and Shobha, 2008). Therefore, using larger values of offset parameter  $d$  in constructing GLCMs for such images will help in extracting more information.

#### 2.4. Haralick texture features

Haralick (Haralick, 1979) proposed the extraction of 14 statistical measures from a GLCM, which can better discriminate the texture. These features found many applications related to texture analysis and pattern classification (Hamilton et al., 2006; Haralick, 1979; Nanni et al., 2010c; Rathore et al., 2014). In this work, we utilized 13 of them including inertia, entropy, energy, correlation, inverse difference moment, sum variance, sum average, sum entropy, difference variance, difference average, difference entropy and two information measures of correlation. As discussed in Section 2.2, feature spaces from  $GLCM_V$  are averaged with  $GLCM_H$  and  $GLCM_D$  with  $GLCM_{oD}$ . The resultant averaged feature spaces are then concatenated with each other to produce 26-D feature space for each protein image.

#### 2.5. The hybrid model

In order to develop efficient classification systems, sometimes different individual feature spaces are combined to collectively utilize their discriminative power. In this study, a number of hybrid models have been developed by concatenating different feature spaces. The detail is given as under:

- $\mathbf{Har}_{24}^{1-5}$  is composed of  $\{\mathbf{Har}_{24}^1, \mathbf{Har}_{24}^2, \mathbf{Har}_{24}^3, \mathbf{Har}_{24}^4, \mathbf{Har}_{24}^5\}$  producing 130-D feature space.
- $\mathbf{Har}_{24}^{6-10}$  is composed of  $\{\mathbf{Har}_{24}^6, \mathbf{Har}_{24}^7, \mathbf{Har}_{24}^8, \mathbf{Har}_{24}^9, \mathbf{Har}_{24}^{10}\}$ . This also produces feature space of 130-D.
- $\mathbf{Har}_{24}^{11-15}$  is composed of  $\{\mathbf{Har}_{24}^{11}, \mathbf{Har}_{24}^{12}, \mathbf{Har}_{24}^{13}, \mathbf{Har}_{24}^{14}, \mathbf{Har}_{24}^{15}\}$ . Here the output feature space dimension is also 130-D.
- $\mathbf{Har}_{24}^{1-15}$  is yielded by the concatenation of  $\mathbf{Har}_{24}^{1-5}, \mathbf{Har}_{24}^{6-10}, \mathbf{Har}_{24}^{11-15}$ . Here 390-D feature space is produced, since this hybrid model is the combination of all the 15 feature spaces for 24 gray levels in the original image.

Such models are developed for all the three datasets.

#### 2.6. Correlation-based Feature Selection

In machine learning, feature selection is a preprocessing technique usually applied on data to select the most informative and discriminative features from the full feature space. In this way, redundant and irrelevant information is removed.

In this paper, we adopted *CFS* feature selection technique that is developed using the concept of filter based feature selection strategy (Hall and Smith, 1999). The concept behind the development of *CFS* is that selected features have high correlation with the target class whereas less correlation with each other. *CFS* utilizes the forward best first search strategy to search the useful features from the full feature space by starting with the empty set. It stops its operations when five consecutive subsets with no improvements are produced.

In this paper, we applied *CFS* on the  $\mathbf{Har}_{24}^{1-15}$  hybrid model for all the three datasets on the intention that first combine the discriminative power of all the feature spaces and then enhance this power by removing the redundant information. In this connection, from 390-D full feature space, *CFS* produced 33-D selected feature space for HeLa dataset, 40-D selected feature space for CHO dataset, and 31-D selected feature space for LOCATE Endogenous dataset.

#### 2.7. Classification and ensemble generation

SVM is a well known and efficient classifier utilized by many researchers for addressing various problems in different application domains (Rathore et al., 2015; Rehman et al., 2013). In the current work, SVM with RBF kernel is trained using  $\mathbf{Har}_g^d$  for  $d = 1, 2, 3, \dots, 15$  and  $g = 4, 8, 12, 16, 20,$  and  $24$ . These classifications will be referred to as  $SVM_g^d$  in this text where  $SVM_g^d$  is trained using  $\mathbf{Har}_g^d$  feature space.

The individual classifications using all the feature spaces are recorded. In order to enhance the prediction performance of the proposed prediction system further, the individual classification results have been combined using Eq. (3).

$$Majority\_Vote = SVM_g^1 * SVM_g^2 * \dots * SVM_g^{15} \quad (3)$$

where  $*$  is the integration operator and *Majority\_Vote* is the final output of the ensemble classifier. While constructing an ensemble, value of  $g$  will be same for all the values of  $d$  as shown in this equation. The complete procedure of *Majority\_Vote* is given as follows. Let us assume the classification results of individual classifiers are  $SVM_g^1, SVM_g^2, \dots, SVM_g^n \in \{CL_1, CL_2, \dots, CL_m\}$ , where  $SVM_g^1, SVM_g^2, \dots, SVM_g^n$  indicate individual base classifiers and  $CL_1, CL_2, \dots, CL_m$  represent labels of protein classes. The output of the ensemble classifier is given as in Eq. (4).

$$Y_i = \sum_{j=1}^n \delta(SVM_g^i, CL_j) \text{ for } j = 1, 2, \dots, n \quad \text{where } n = 15 \quad (4)$$

where

$$\delta(SVM_g^i, CL_j) = \begin{cases} 1 & \text{if } SVM_g^i \in CL_j \\ 0 & \text{otherwise} \end{cases}$$

The final output is obtained through the integration of individual prediction results using the majority voting scheme as given in Eq. (5).

$$Y_{Majority\_Vote} = \max \{Y_1, Y_2, \dots, Y_n\} \quad (5)$$

where  $Y_{Majority\_Vote}$  is the output prediction of the ensemble.

Ensembles are built corresponding to all the offset values of every gray level quantization. The best ensemble is selected for building the final model of *GLCM-SubLoc* protein classification system.

### 3. Results and discussion

Results are presented and analyzed in this section. Accuracy (Acc), sensitivity (Sen), specificity (Spe), MCC, F-score, and Q statistics are employed to evaluate the performance of the proposed *GLCM-SubLoc* prediction system. We utilized 10-fold cross validation protocol to validate the performance of the system.

#### 3.1. Performance analysis of *GLCM-SubLoc* for HeLa dataset

The performance accuracies of the proposed *GLCM-SubLoc* prediction system are shown in Table 1. The first column indicates the value of the parameter  $d$ , which is the distance

**Table 1** Performance of individual SVMs using  $\text{Har}_g^d$  features for different  $d$  and  $g$  values utilizing HeLa dataset.

$d$	$G$					
	4	8	12	16	20	24
1	69.3	75.7	78.1	79.5	80.8	80.9
2	70.1	72.7	75.6	78.3	81.2	80.9
3	69.2	72.1	75.9	78.8	80.8	80.7
4	69.8	71.5	76.1	78.3	80.2	80.2
5	66.9	73.2	75.8	78.3	80.9	80.9
6	68.2	71.9	74.8	76.6	78.8	79.8
7	66.2	70.8	74.8	76.7	79.9	79.8
8	66.0	71.1	74.7	76.9	79.3	79.9
9	65.3	71.1	74.8	76.9	79.6	80.3
10	67.4	71.5	74.3	77.9	81.2	81.0
11	65.6	71.4	75.2	78.6	80.5	80.8
12	67.0	70.9	73.8	77.8	80.2	80.6
13	67.7	71.1	73.8	76.7	79.4	79.9
14	66.0	71.2	75.0	77.6	79.5	81.3
15	66.0	71.4	75.9	77.4	80.1	81.2

between the pixels to be checked for co-occurrence in constructing GLCM. The rest of the columns show the accuracy values yielded by *GLCM-SubLoc* utilizing the features from  $\text{GLCM}_g^d$  for different  $d$  and  $g$  values.

It is observed that using the same gray level for different values of distance  $d$  varying from 1 to 15, there is very little variation in the accuracy values. The best values are obtained for  $d = 1$  in all cases except for the gray level value of 24 where the best value of accuracy is yielded for  $d = 14$ . On the other hand, gradual enhancement in the accuracy values for the same value of  $d$  with varying gray levels is observed. The difference is initially higher, however, for gray levels 16, 20 and 24, this difference is small.

Almost same performance by the prediction system for different values of  $d$  shows that the patterns and information present in HeLa protein images is identical. The co-occurrence distribution of different intensities over the protein image is leading the classifier to show the same performance. However, when more information is added to the GLCM by varying the gray levels, the prediction performance of proposed model is enhanced. For the gray levels 20 and 24, the prediction performance has not been improved due to the redundancy in the information preserved by the two GLCMs from which the features are extracted.

**Supplementary Table 1** demonstrates the detailed results for HeLa dataset. The performance of the proposed prediction system is demonstrated through sensitivity, specificity, MCC and F-score for different gray levels and for different values of GLCM offset parameter. Quite good performance is revealed from the quantitative analysis of the presented predic-

tion values. The performance measures validate the good performance of the proposed prediction system.

The performance of the proposed prediction system using different hybrid models is shown in **Table 2**. The hybrid models have shown improved performance in some cases whereas comparable performance in some other.

The  $\text{Har}_{24}^{1-5}$  hybrid model has shown 5.6% improvement over the individual features for gray level value  $g = 24$  and  $d$  values 1–5 where the highest accuracy in the range from 1 to 5 in **Table 1** is 80.9%. However,  $\text{Har}_{24}^{6-10}$  hybrid model has shown low accuracy compared to the accuracies in the range of 6–10 for  $d$  values against gray level value  $g = 24$  where the highest accuracy in this range is 81.0%. Similarly, the performance of  $\text{Har}_{24}^{11-15}$  hybrid model is lower than the individual accuracies in the range 11–15 for different  $d$  values for the same gray level value. On the other hand, the performance of the proposed prediction system using  $\text{Har}_{24}^{1-15}$  hybrid model is 88.2%, which is 6.9% higher than the highest accuracy in the range 1–15 for different  $d$  values. This reveals that the hybrid model has enhanced the prediction performance of the proposed prediction system compared to the individual feature spaces. The performance of the proposed prediction system using the CFS based selected feature space has shown comparable performance that is 88.3% compared to the  $\text{Har}_{24}^{1-15}$  hybrid model. However, the dimensionality of the CFS based features is 33-D compared to the 390-D dimensionality of the full feature space. The results obtained through the majority voting scheme for HeLa dataset are presented in **Table 3**. For each gray level value under consideration there are

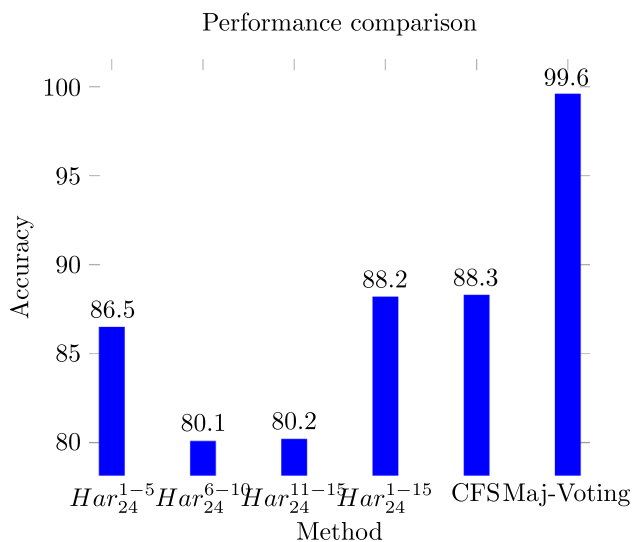
**Table 2** Performance of the proposed prediction system using the hybrid models for HeLa dataset.

Features	Acc	Sen	Spe	MCC	F-score
$\text{Har}_{24}^{1-5}$	86.5	89.6	86.1	0.56	0.58
$\text{Har}_{24}^{6-10}$	80.1	84.2	79.6	0.44	0.46
$\text{Har}_{24}^{11-15}$	80.2	84.5	79.7	0.44	0.47
$\text{Har}_{24}^{1-15}$	<b>88.2</b>	90.5	87.9	0.59	0.61

The bold values show the highest achieved accuracy.

**Table 3** Ensemble performance of *GLCM-SubLoc* for HeLa dataset.

	<i>G</i>					
	4	8	12	16	20	24
Acc	91.0	96.8	98.9	99.6	99.5	99.6
<i>Q</i> -value	0.33	0.25	0.21	0.21	0.24	0.25



**Figure 4** Performance comparison for different models of HeLa dataset.

GLCMs for different values of offset parameter  $d$  ranging from 1 to 15 as discussed earlier. Ensemble is constructed from the results of 15 classifications for the same gray level. For example, taking into account GLCM of 4 gray levels, there are 15 values for the parameter  $d$  hence producing 15 classifications. In order to form an ensemble, the 15 classifications for the GLCM of 4 gray levels are employed.

The highest ensemble accuracy is reported for the ensemble constructed for the gray levels 16 and 24. However, the similarity measure  $Q$ -statistic shows low similarity among the members of the ensemble using GLCMs of 16 gray levels. This indicates that GLCMs of 16 gray levels produce informative and discriminative features for the classification stage. Fig. 4 shows the comparison among the  $Har_{24}^{1-5}$ ,  $Har_{24}^{6-10}$ ,  $Har_{24}^{11-15}$ ,  $Har_{24}^{1-15}$ , the CFS based selected feature space, and the results of the majority voting based ensemble.

The highest performance among different approaches for HeLa dataset is achieved through the majority voting scheme, which is yielded by integrating the decisions of different support vector machines.

### 3.2. Performance analysis of *GLCM-SubLoc* prediction system for CHO dataset

Performance assessments of the proposed prediction system for CHO dataset are presented in Table 4. The first column shows different values for the GLCM offset parameter  $d$  whereas the rest of the columns show accuracy values for the  $Har_g^d$  features extracted from  $GLCM_g^d$ .

It is observed that for a certain gray level, the accuracy value keeps increasing for the higher value of GLCM offset parameter  $d$ . This indicates that the co-occurring intensity values at larger distances possess more discriminative information compared to smaller values of  $d$  for this particular dataset. That is why the discriminative power of the classification system is enhanced with larger distances. Similarly, for a certain

**Table 4** Performance of individual support vector machines using  $Har_g^d$  features for different  $d$  and  $g$  values utilizing CHO dataset.

$d$	$g$					
	4	8	12	16	20	24
1	73.9	80.0	84.5	85.7	83.8	83.6
2	75.5	80.6	85.0	85.4	85.6	86.0
3	77.8	80.5	86.5	86.0	85.9	86.2
4	79.7	82.7	86.6	86.0	87.8	87.4
5	79.7	83.8	87.1	87.4	87.2	87.8
6	80.8	85.1	86.5	87.4	87.8	88.0
7	80.8	84.1	86.9	88.0	88.0	87.8
8	81.4	86.2	87.1	88.4	88.1	89.2
9	81.2	85.4	87.4	88.1	87.8	88.9
10	81.1	86.5	88.3	88.3	88.3	<b>89.3</b>
11	81.5	85.3	87.1	88.1	88.1	88.7
12	82.0	85.4	88.0	88.0	88.7	89.0
13	82.0	85.1	87.4	87.8	88.9	88.9
14	82.1	85.3	87.8	87.8	89.0	89.0
15	81.5	85.6	87.7	88.1	89.3	<b>89.3</b>

The bold values show the highest achieved accuracy.

**Table 5** Performance of the proposed prediction system using the hybrid models for CHO dataset.

Features	Acc	Sen	Spe	MCC	F-score
$\text{Har}_{24}^{1-5}$	88.9	88.3	89.0	0.68	0.73
$\text{Har}_{24}^{6-10}$	89.2	89.6	89.0	0.69	0.74
$\text{Har}_{24}^{11-15}$	89.2	89.3	89.2	0.69	0.74
$\text{Har}_{24}^{1-15}$	<b>90.7</b>	91.3	90.5	0.73	0.77

The bold values show the highest achieved accuracy.

**Table 6** Ensemble performance of *GLCM-SubLoc* for CHO dataset.

	Gray levels					
	4	8	12	16	20	24
Acc	94.6	98.6	<b>100</b>	<b>100</b>	99.8	99.8
$Q$ -value	0.34	0.21	0.12	0.12	0.19	0.20

The bold values show the highest achieved accuracy.

value of the distance parameter  $d$ , higher accuracies are achieved while utilizing features from GLCMs with more gray levels. In these results, GLCMs with gray level value of 16 might be enough for feature extraction.

This shows that prior to the development of a prediction model for classification, protein images coming from a particular source should be analyzed first for their patterns. Here, it is observed that higher accuracies are achieved for CHO dataset when the value of GLCM distance parameter  $d$  is 8 or above for all the gray levels presented in Table 4. Detailed results for CHO dataset are shown in Supplementary Table 2 where the performance achieved is demonstrated in terms of sensitivity, specificity, MCC and F-score. The assessments shown with different measures validate the performance of the proposed prediction model.

Hybrid models, utilized by the proposed prediction system, have shown improvement over the individual feature spaces as shown in Table 5.  $\text{Har}_{24}^{1-15}$  achieved 1.4% higher accuracy compared to the highest accuracy achieved by the same features for  $d = 15$  and  $g = 24$  as shown in Table 4.

However, CFS based feature selection has also shown comparable performance against the hybrid models. This shows that feature selection or hybridization is unable to improve the performance of the prediction system. However, CFS based feature selection has reduced the dimensionality of the feature space from 390-D to 40-D only. On the other hand, the majority voting based ensemble outperformed all the adopted approaches in predicting protein localization from CHO dataset.

Table 6 presents the ensemble performance for CHO dataset. The first row shows the ensemble accuracy whereas the second row indicates the value for  $Q$ -statistic.

The highest ensemble accuracy is reported for the GLCMs of 12 and 16 gray levels where the accuracy value is 100% in both cases. The  $Q$ -statistic value shows the highest diversity for the members of these two ensembles. The GLCMs of gray level 12 and 16 for different values of distance parameter  $d$  produce feature spaces with diversified information, hence producing classifications of higher diversities. This leads to the higher performance of the ensemble classifier.

Performance comparison

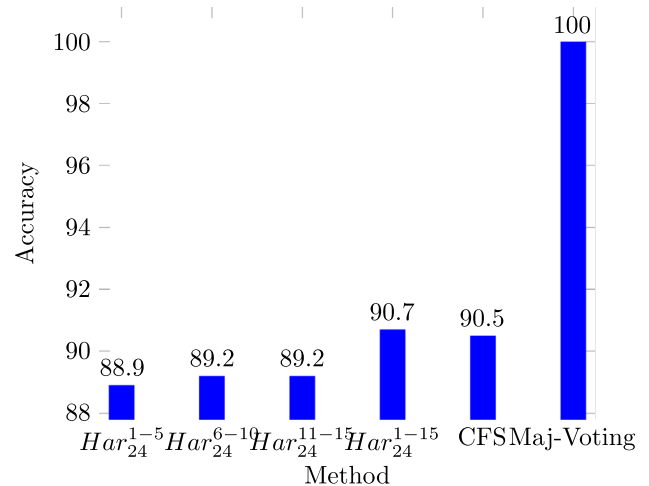
**Figure 5** Performance comparison for different models of CHO dataset.

Fig. 5 shows the comparison among the  $\text{Har}_{24}^{1-5}$ ,  $\text{Har}_{24}^{6-10}$ ,  $\text{Har}_{24}^{11-15}$ ,  $\text{Har}_{24}^{1-15}$ , the CFS based selected feature space, and the results of the majority voting based ensemble.

It is evident from Fig. 5 that majority voting based ensemble has shown significant performance in classifying protein images from fluorescence microscopy images of CHO dataset.

### 3.3. Performance analysis of *GLCM-SubLoc* prediction system for LOCATE Endogenous dataset

Performance accuracies of the proposed *GLCM-SubLoc* prediction system are presented in Table 7 for LOCATE Endogenous dataset. The first column shows different values of distance parameter  $d$  for GLCM construction whereas the remaining columns show the accuracy values achieved using features extracted from GLCM with different gray levels varying from 4 to 24.



**Table 7** Performance of individual support vector machines using  $\text{Har}_g^d$  features for different  $d$  and  $g$  values utilizing LOCATE Endogenous dataset.

$d$	$g$					
	4	8	12	16	20	24
1	79.0	84.8	86.8	88.6	88.2	88.2
2	82.0	86.4	87.8	88.2	88.4	88.8
3	82.0	85.8	88.4	<b>89.2</b>	87.8	88.4
4	79.4	87.6	87.6	88.4	86.8	87.2
5	77.6	86.8	86.8	87.2	85.8	87.0
6	78.4	87.2	86.8	87.2	86.0	86.6
7	76.8	87.2	86.8	87.2	85.4	87.2
8	78.0	87.2	86.4	86.2	85.2	86.2
9	78.8	85.8	85.4	84.4	85.2	86.4
10	80.6	85.8	85.6	84.6	84.0	85.8
11	80.4	85.2	87.6	85.2	85.0	87.2
12	80.6	86.0	85.6	84.6	84.6	85.0
13	80.8	86.0	86.2	85.2	85.4	86.2
14	80.4	86.0	85.6	86.2	85.0	85.2
15	79.4	86.2	86.6	86.2	86.2	86.4

The bold values show the highest achieved accuracy.

**Table 8** Performance of the proposed prediction system using the hybrid models for LOCATE Endogenous dataset.

Features	Acc	Sen	Spe	MCC	F-score
$\text{Har}_{24}^{1-5}$	90.6	91.5	90.5	0.64	0.65
$\text{Har}_{24}^{6-10}$	88.2	89.2	88.1	0.58	0.59
$\text{Har}_{24}^{11-15}$	87.2	88.5	87.1	0.56	0.57
$\text{Har}_{24}^{1-15}$	<b>92.6</b>	92.0	92.6	0.69	0.71

The bold values show the highest achieved accuracy.

**Table 9** Ensemble performance of *GLCM-SubLoc* for LOCATE Endogenous dataset.

	Gray levels					
	4	8	12	16	20	24
Acc	99.8	100	99.8	99.6	100	100
$Q$ -value	0.20	0.18	0.23	0.26	0.20	0.14

The results presented show that there is some improvement in the classification accuracy for smaller values of  $d$ . For example, performance is better using the  $\text{Har}_4^3$ ,  $\text{Har}_{12}^3$ , and  $\text{Har}_{16}^3$  features from  $\text{GLCM}_4^3$ ,  $\text{GLCM}_{12}^3$ , and  $\text{GLCM}_{16}^3$ , respectively. The highest accuracy values using gray level values of 4, 12, and 16 with  $d = 3$  are 82.0%, 88.4%, and 89.2%, respectively. Further increasing the value of  $d$  for a certain gray level did not enhance the performance. This shows that for LOCATE Endogenous dataset, a reasonable value of  $d$  might be between 2 and 8 inclusive. For other values, the performance is usually observed to be degraded.

As far as gray levels are concerned, 16 gray levels are sufficient to generate informative features from respective GLCMs for protein images of LOCATE Endogenous dataset as can be observed from the results shown in Table 7.

The detailed performance predictions are provided in Supplementary Table 3, which demonstrates the performance of

the prediction system in terms of sensitivity, specificity, MCC and F-score. The performance predictions show that the proposed prediction system has efficiently classified the protein images from LOCATE Endogenous dataset into different classes.

The hybrid models for LOCATE Endogenous dataset, as shown in Table 8, have shown improvement over individual feature spaces. The highest performance accuracy 92.6% is achieved by the proposed prediction system using  $\text{Har}_{24}^{1-15}$  hybrid model. This signifies the collective discrimination power of all the individual feature spaces.

Similarly, the CFS based selected feature space has shown slight improvement over  $\text{Har}_{24}^{1-15}$  hybrid model, which is good achievement as far as dimensionality of the feature space is concerned. The CFS based feature space dimensionality is only 31-D achieving 92.2% accuracy as compared to the 390-D dimensionality of the full feature space achieving 92.6% accuracy.

The ensemble performance for LOCATE Endogenous dataset is shown in Table 9. The highest ensemble accuracies are recorded for GLCMs of 8, 20 and 24 gray levels. However, the highest diversity is achieved for the members of ensemble at gray levels 24 as shown by the  $Q$ -statistic value of 0.14.

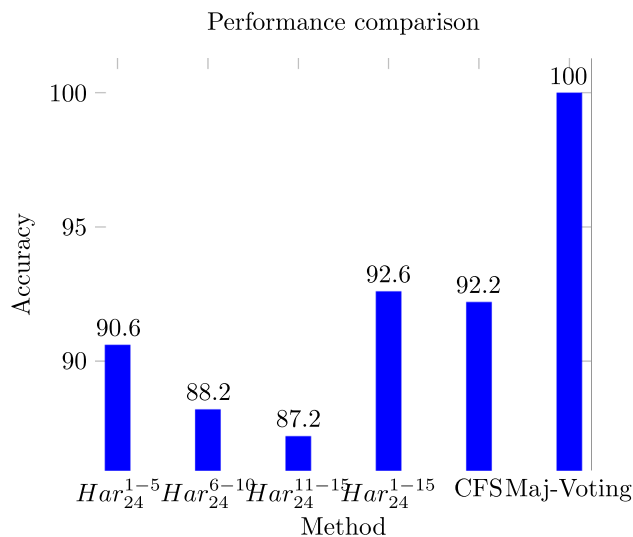
This indicates that for this particular dataset, GLCMs might be constructed with either 8, 20 or 24 gray levels. The comparison among  $Har_{24}^{1-5}$ ,  $Har_{24}^{6-10}$ ,  $Har_{24}^{11-15}$ ,  $Har_{24}^{1-15}$ , the CFS based selected feature space, and the results of the majority voting based ensemble is shown in Fig. 6.

The performance comparison for LOCATE Endogenous dataset reveals that majority voting based ensemble outperformed other approaches.

#### 4. Comparative analysis

Comparison of the proposed prediction system with state-of-the-art techniques is shown in Table 10. Chebira et al. (2007) have proposed a prediction system that utilized features from multi-resolution subspaces. This system achieved 95.4% prediction accuracy for HeLa dataset. Lin et al. (2007) have proposed a variant of AdaBoost learning algorithm that yielded 93.6% and 94.7% prediction accuracies for HeLa and CHO datasets, respectively. Nanni et al. (2010c) developed a prediction system, based on the fusion of two ensembles, which yielded 97.5% performance accuracy for HeLa dataset. In another approach, Nanni et al. (2010a) have identified some novel features in conjunction with random subspace of Neural Networks for classifying protein images from HeLa and LOCATE datasets. Their prediction system achieved 95.8% and 99.5% accuracies for HeLa and LOCATE Endogenous datasets, respectively.

Similarly, Nanni et al. (2010b) have developed an ensemble of SVMs trained on a random subset of features extracted from binary and ternary patterns. Their prediction system obtained 93.2% performance accuracy for HeLa dataset whereas 92.9% performance accuracy for LOCATE Endogenous dataset.



**Figure 6** Performance comparison for different models of LOCATE Endogenous dataset.

**Table 10** Performance comparison with other published work.

Method	Accuracy		
	HeLa	CHO	LOCATE
Chebira et al. (2007)	95.4	–	–
Lin et al. (2007)	93.6	94.7	–
Nanni et al. (2010c)	97.5	–	–
Nanni et al. (2010a)	95.8	–	99.5
Nanni et al. (2010b)	93.2	–	92.9
Tahir et al. (2012)	99.7	–	99.8
Tahir et al. (2013)	–	96.5	–
Tahir et al. (2014)	100	95	–
GLCM-SubLoc	99.65	100	100

Furthermore, Tahir et al. have shown the importance of feature extraction in multi-resolution subspaces in conjunction with majority voting based ensemble (Tahir et al., 2012), which achieved 99.7% and 99.8% prediction accuracies for HeLa and LOCATE Endogenous datasets, respectively. In another approach, Tahir et al. have shown empirically that introducing synthetic samples in the feature space increases the classifier's bias toward the minority class and ultimately enhances the prediction accuracy (Tahir et al., 2013). This system achieved 96.5% accuracy for CHO dataset. Similarly, another approach proposed by Tahir et al. (2014) has achieved 100% prediction accuracy for HeLa and 95.0% accuracy for CHO dataset.

The prediction system proposed in this work has achieved 99.6% prediction accuracy for HeLa dataset and 100% accuracy for each of the CHO and LOCATE Endogenous datasets.

#### 5. Conclusive remarks

In this paper, we performed extensive empirical analysis of fluorescence microscopy protein images using GLCM based textural features. We combined different approaches as well as adapted feature selection strategy to remove redundant and irrelevant information. The proposed method was further enhanced by applying the majority voting based technique in decision making. We showed that considering different values of the offset parameter  $d$  in the construction of GLCM plays a key role in the extraction of diversified information from fluorescence microscopy protein images. Similarly, we provided empirical evidence that quantization levels up to 24 gray level values is sufficient to transform an image into an informative GLCM.

We also utilized correlation based feature selection in order to extract most useful information from the full feature space as well as to reduce the dimensionality. The simulation results showed that performance of the proposed prediction system was consistent with reduced feature spaces. We utilized three benchmark protein image datasets in order to validate the significance of *GLCM-SubLoc* prediction system, hence providing empirical evidence on the generalization capability of the proposed system. The results are compared against state-of-the-art approaches.

As far as our point of view is concerned, the most precious finding of this work is successful demonstration of the effec-

tiveness of the features extracted from GLCM while considering different offset values and various quantization levels. Thus opening new avenues for other researches in the field to explore novel techniques for extracting information from GLCM matrices.

### Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jksus.2016.12.004>.

### References

- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., 2002. *Molecular Biology of the Cell*. National Center for Biotechnology Information~Os Bookshelf.
- Boland, M.V., Markey, M.K., Murphy, R.F., 1998. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry* 33, 366–375.
- 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.
- Chaddad, A., Tanougast, C., Dandache, A., Bouridane, A., 2011. Improving of colon cancer cells detection based on Haralick's features on segmented histopathological images. In: *IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, 2011. IEEE, Penang, pp. 87–90.
- Chebra, 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 Bioinf.* 8, 210.
- Chen, S., Wu, C., Chen, D., Tan, W., 2009. Scene classification based on gray level-gradient co-occurrence matrix in the neighborhood of interest points. In: *IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2009. ICIS 2009. IEEE, Shanghai, pp. 482–485.
- Chen, X., Velliste, M., Murphy, R.F., 2006. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. *Cytom. Part A J. Int. Soc. Adv. Cytom.* 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. *J. Theor. Biol.* 318, 1–12.
- Cooper, G.M., 2000. *The Cell: A Molecular Approach*, second ed. ASM Press, Sunderland, Mass.: Sinauer Associates, Washington, D.C.
- Do-Hong, T., Le-Tien, T., Bui-Thu, C., 2010. A new descriptor for image retrieval using contourlet co-occurrence. In: *Third International Conference on Communications and Electronics (ICCE)*, 2010. IEEE, Nha Trang, pp. 169–174.
- Gelzinis, A., Verikas, A., Bacauskiene, M., 2007. Increasing the discrimination power of the co-occurrence matrix-based features. *Pattern Recognit.* 40, 2367–2372.
- Glory, E., Murphy, R.F., 2007. Automated subcellular location determination and high-throughput microscopy. *Dev. Cell* 12, 7–16.
- Hall, M.A., Smith, L.A., 1999. Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. In: *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*. AAAI Press, Orlando, Florida, pp. 235–239.
- Hamilton, N.A., Pantelic, R.S., Hanson, K., Fink, J.L., Karunarathne, S., Teasdale, R.D., 2006. Automated subcellular phenotype classification: an introduction and recent results. 2006 Work. Intell. Syst. Bioinforma. (WISB 2006).
- Haralick, R.M., 1979. Statistical and structural approaches to texture. *Proc. IEEE*. <http://dx.doi.org/10.1109/PROC.1979.11328>.
- Li, J., Xiong, L., Schneider, J., Murphy, R.F., 2012. Protein subcellular location pattern classification in cellular images using latent discriminative models. *Bioinformatics* 28, i32–i39.
- Liang, M., Malm, P., 2012. 3D Co-occurrence Matrix Based Texture Analysis Applied to Cervical Cancer Screening. *Dep. Inf. Technol.* Uppsala University, Uppsala, Sweden.
- Lin, C.-C., Tsai, Y.-S., Lin, Y.-S., Chiu, T.-Y., Hsiung, C.-C., Lee, M.-L., Simpson, J.C., Hsu, C.-N., 2007. Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization. *Bioinformatics* 23, 3374–3381.
- Mitreă, D., Mitreă, P., Nedevschi, S., Badea, R., Lupsor, M., Hagiu, C., Ciobanu, L., 2012. Abdominal tumor characterization and recognition using superior-order cooccurrence matrices, based on ultrasound images. *Comput. Math. Methods Med.* 2012, 17.
- Murphy, R.F., Velliste, M., Porreca, G., 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., Brahman, S., Ghidoni, S., Menegatti, E., Barrier, T., 2013. Different approaches for extracting information from the co-occurrence matrix. *PLoS One* 8, e83554.
- Nanni, L., Brahnam, S., Lumini, A., 2010a. Novel features for automated cell phenotype image classification. *Adv. Comput. Biol. Adv. Exp. Med. Biol.* 680, 207–213.
- Nanni, L., Brahnam, S., Lumini, A., 2010b. Selecting the best performing rotation invariant patterns in local binary/ternary patterns. *Proc. Int. Conf. Image Process. Comput. Vision Pattern Recognit.*
- Nanni, L., Lumini, A., Lin, Y.-S., Hsu, C.-N., Lin, C.-C., 2010c. Fusion of systems for automated cell phenotype image classification. *Expert Syst. Appl.* 37, 1556–1562.
- Newberg, J., Hua, J., Murphy, R.F., 2009. Location proteomics: systematic determination of protein subcellular location. In: Maly, I.V. (Ed.), *Systems Biology*. Humana Press, pp. 313–332.
- Rathore, S., Hussain, M., Iftikhar, M.A., Jalil, A., 2015. Novel structural descriptors for automated colon cancer detection and grading. *Comput. Methods Programs Biomed.* 121, 92–108.
- Rathore, S., Hussain, M., Iftikhar, M.A., Jalil, A., 2014. Ensemble classification of colon biopsy images based on information rich hybrid features. *Comput. Biol. Med.* 47, 76–92.
- Rehman, Z.U., Mirza, M.T., Khan, A., Xhaard, H., 2013. Chapter four – Predicting G-protein-coupled receptors families using different physiochemical properties and pseudo amino acid composition. *Methods Enzymol.* 522, 61–79.
- Srinivasa, G., Merryman, T., Chebra, A., Kovacevic, J., Mintos, A., 2006. Adaptive multiresolution techniques for subcellular protein location classification. *IEEE Int. Conf. Acoust. Speech Signal Process.*
- Srinivasan, G.N., Shobha, G., 2008. Statistical texture analysis. *Proc. World Acad. Sci. Eng. Technol.* 36, 1264–1269.

- Tahir, M., Khan, A., Kaya, H., 2014. Protein subcellular localization in human and hamster cell lines: employing local ternary patterns of fluorescence microscopy images. *J. Theor. Biol.* 340, 85–95.
- Tahir, M., Khan, A., Majid, A., 2012. Protein subcellular localization of fluorescence imagery using spatial and transform domain features. *Bioinformatics* 28, 91–97. <http://dx.doi.org/10.1093/bioinformatics/btr624>.
- Tahir, M., Khan, A., Majid, A., Lumini, A., 2013. Subcellular localization using fluorescence imagery: utilizing ensemble classification with diverse feature extraction strategies and data balancing. *Appl. Soft Comput.* 13, 4231–4243.
- Tscherepanow, M., Jensen, N., Kummert, F., 2008. An incremental approach to automated protein localisation. *BMC Bioinf.* 9, 445.
- Walker, R.F., Jackway, P.T., Longstaff, D., 2003. Genetic algorithm optimization of adaptive multi-scale GLCM features. *Int. J. Pattern Recognit. Artif. Intell.* 17, 17–39.