Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Automated colon cancer detection using hybrid of novel geometric features and some traditional features

Saima Rathore [a,b,*], Mutawarra Hussain [a], Asifullah Khan [a]

[a] DCIS, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan
[b] DCS&IT, University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir

ARTICLE INFO

Article history:
Received 31 October 2014
Accepted 6 March 2015

Keywords:
Colon cancer
Classification
Elliptic objects
Colon biopsy
Elliptic Fourier descriptors
SIFT
Morphological, geometric, and texture features

ABSTRACT

Automatic classification of colon into normal and malignant classes is complex due to numerous factors including similar colors in different biological constituents of histopathological imagery. Therefore, such techniques, which exploit the textural and geometric properties of constituents of colon tissues, are desired. In this paper, a novel feature extraction strategy that mathematically models the geometric characteristics of constituents of colon tissues is proposed. In this study, we also show that the hybrid feature space encompassing diverse knowledge about the tissues' characteristics is quite promising for classification of colon biopsy images. This paper thus presents a hybrid feature space based colon classification (HFS-CC) technique, which utilizes hybrid features for differentiating normal and malignant colon samples. The hybrid feature space is formed to provide the classifier different types of discriminative features such as features having rich information about geometric structure and image texture. Along with the proposed geometric features, a few conventional features such as morphological, texture, scale invariant feature transform (SIFT), and elliptic Fourier descriptors (EFDs) are also used to develop a hybrid feature set. The SIFT features are reduced using minimum redundancy and maximum relevancy (mRMR). Various kernels of support vector machines (SVM) are employed as classifiers, and their performance is analyzed on 174 colon biopsy images. The proposed geometric features have achieved an accuracy of 92.62%, thereby showing their effectiveness. Moreover, the proposed HFS-CC technique achieves 98.07% testing and 99.18% training accuracy. The better performance of HFS-CC is largely due to the discerning ability of the proposed geometric features and the developed hybrid feature space.

1. Introduction

Medical imaging has gained much importance in the last few decades, especially in analyzing different body parts for predicting certain disorders/diseases. Microscopic imaging is one of the medical imaging techniques, wherein the images of biopsy slides are captured. Biopsy images have well-defined organization of tissues and connected components, depending upon the body part from which they are taken [1]. The same is true for colon biopsy images, which are used in our problem for cancer detection. Biologically different constituents in a colon biopsy image can be identified by looking at the spatial organization of its constituents.

Microscopic analysis is the commonly practiced technique of colon cancer diagnosis, wherein histopathologists visually examine the deformation of tissues under microscope, and decide whether the geometric structure and organizational arrangement of sample tissues belong to the class of malignant or normal colon. Microscopic analysis is time consuming as well as subjective. The main reason behind subjectivity is the fact that quantitative cancer grades are assigned depending upon the observed morphology of tissues by the histopathologists. This process also leads to inter/intra-observer variability as quantitative grades assigned to the same sample by different histopathologists, or even by one histopathologist, may vary at times [2,3]. In order to alleviate such problems in diagnosis, researchers are working since long to find automatic quantitative tools, which could measure the degree of deformation and assign quantitative cancer grades to the colon samples.

The research in the field of colon cancer is in various dimensions. A larger subset of the colon cancer detection techniques has been summarized in a recent survey reported by Rathore et al. [4]. Some authors have performed analysis on hyperspectral data of colon biopsies [5,6]. In these schemes, authors select one spectral band amongst several bands of hyperspectral cube, calculate image features, and then based on these features classify samples

* Corresponding author at: DCIS, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan. Tel.: +92 51 2207381 × 3102.
E-mail address: saimarathore_2k6@yahoo.com (S. Rathore).

http://dx.doi.org/10.1016/j.compbiomed.2015.03.004
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

into multiple classes. Some researchers have studied thousands of human genes in parallel by using two variants of microarrays [7,8]. Their aim was to identify such genetic alterations, which were supposed to be responsible for colon cancer. Like genes, blood serum also deviates from its normal composition in case of colon cancer. Researchers have exploited this variation, and have used laser-induced fluorescence and Raman spectroscopy of blood serum for cancer detection [9,10].

Some researchers have exploited the variation in the texture of normal and malignant colon biopsy images for cancer detection. In this context, Esgiar et al. analyzed distinctiveness of six texture features (angular second moment, contrast, correlation, entropy, inverse difference moment, and dissimilarity) for classification of colon biopsy images [11]. They found the combination of entropy and correlation to be the most distinctive feature set, providing an overall accuracy of 90.2%. They further extended the idea by introducing fractal dimensions into the classification process, and proved that a combination of entropy, correlation and image fractal dimensions yields classification accuracy of 94.1% [12]. Followed by their work, Masood et al. proposed a few valuable methodologies for classification of colon. In their first method, they calculated morphological features of shape, size and orientation, and gray-level co-occurrence matrix (GLCM) based features of energy, inertia, and local homogeneity from colon biopsy images [13]. They employed polynomial SVM classifier, and achieved classification accuracy of 84% and 90% using morphological and GLCM based features, respectively. Masood et al. further extended the previous work [13], and calculated circular local binary patterns in order to classify colon biopsy images [14]. They obtained an accuracy of 90% by employing Gaussian SVM for classification. Further, Rathore et al. proposed a colon biopsy image based classification technique (CBIC) [15], wherein a hybrid feature set comprising traditional histogram of oriented gradients based features, and novel variants of statistical moments and Haralick texture features has been used for classification of colon biopsy images. A majority voting based ensemble of SVM classifiers has been used for classification, and 98.85% classification accuracy has been observed.

Recently, Altunbay et al. proposed a colon cancer detection technique [16], wherein they constructed a graph on different objects, obtained by using circle fitting algorithm [1] on the white, pink and purple clusters of colon biopsy image. Features of degree, average clustering coefficient, and diameter are computed from the graphs. The features are then used to classify given samples by using linear SVM kernel. In addition, Ozdemir et al. presented a method for automated colon cancer detection [17]. In this work, reference graphs of a few normal images are generated by employing previously proposed method of graph creation [18], and are stored for further referencing. Some query graphs are generated from the test images, and are searched in the reference graphs. Three most similar graphs are found in the reference images. Finally, normal or malignant class is assigned to the test sample based on the degree of similarity of the query graph with the three most similar graphs.

Table 1
Abbreviations used in the text.

Acronym	Abbreviations
EFDs	Elliptic Fourier descriptors
FEM	Feature extraction module
H&E	Hematoxylin & Eosin
MCC	Matthews correlation coefficient
mRMR	Minimum redundancy and maximum relevancy
OSDU	Object spatial distribution uniformity
OSU	Object size uniformity
RBF	Radial basis function
SIFT	Scale invariant feature transform
SVM	Support vector machine

The techniques mentioned in the previous paragraphs have a few limitations. First, graph based techniques [16,17] are computationally expensive. Second, texture features based techniques [11–15] have exploited general texture features for classification, and have not exploited the background knowledge about the morphology of colon tissues for classification. Therefore, a computer-aided diagnostic technique, which could exploit the morphology of normal and malignant colon tissues in a computationally tractable manner, is required.

In this research, a novel hybrid feature space based colon classification (HFS-CC) technique has been proposed for robust and effective classification of colon biopsy images. We propose a novel feature type that mathematically quantifies the geometric structure of constituents of colon tissues. Further, we compute some other feature types such as morphological, SIFT, EFDs, and texture features, and combine those features with geometric features to form a hybrid feature set. HFS-CC differs from its previous counterparts in two aspects. First, it incorporates background knowledge about tissues organization into the classification process by introducing novel geometric features, thus leads to effective and promising results. Second, it caters different categories of features, and exploits positive aspects of each category. There are several abbreviations used in subsequent sections. These abbreviations are given in Table 1.

The remainder of this paper is organized as follows. Section 2 describes the structure of normal and malignant colon tissues. Section 3 presents the proposed scheme in detail. Section 4 describes performance measures. Section 5 demonstrates experimental results, and Section 6 concludes the paper.

2. Cell structure: Normal and malignant colon tissues

Normal colon tissues have well-defined organizational structure. There are three constituents of a normal colon tissue, namely, epithelial cells, non-epithelial cells, and connecting tissues. The detailed structure of a normal colon tissue is shown in Fig. 1(a). Lumen lies in the middle of the tissue and is surrounded by epithelial cells that form glandular structure, whereas, non-epithelial cells lie in between glandular structures and are known as stroma. Cells and connected tissues are organized and coherent in case of normal colon. But, this organizational structure deviates considerably for malignant tissues as shown in Fig. 1(b)–(d).

Histopathologists analyze the samples under microscope and decide whether tissue is normal or not. Furthermore, histopathologists also assign quantitative cancer grades to the malignant colon samples. Grade of colon cancer is the differentiability level of malignant tissues from normal ones. There are three colon cancer grades: well-, moderately-, and poorly differentiable. In a well differentiable grade, tissues are slightly similar to normal ones as shown in Fig. 1(b). In this particular grade, cancer progresses at low speed. In moderately differentiable cancer grade, tissues are different from normal ones as shown in Fig. 1(c), and cancer progresses at moderate speed in this grade. In a poorly differentiable cancer grade, malignant tissues are totally different from normal tissues as shown in Fig. 1(d), and cancer progresses at very high rate in this particular grade.

3. Proposed scheme

The proposed HSF-CC scheme comprises six main stages, namely, preprocessing, feature extraction, feature reduction, feature concatenation,

Please cite this article as: S. Rathore, et al., Automated colon cancer detection using hybrid of novel geometric features and some traditional features, Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.03.004

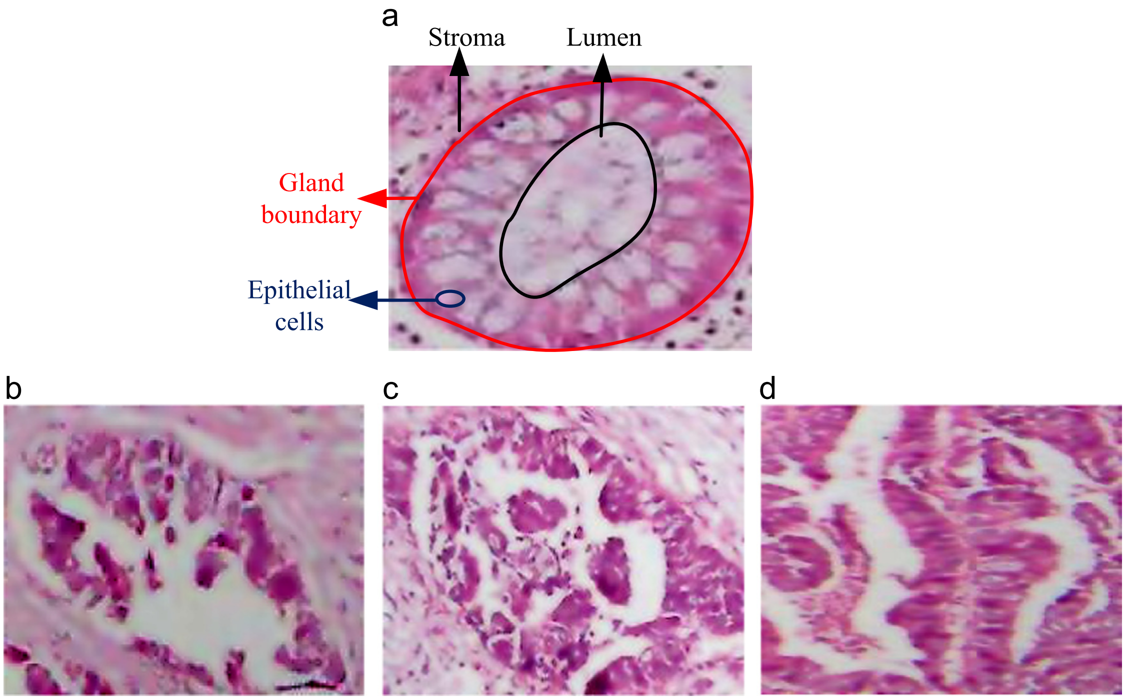


Fig. 1. Organizational structure of (a) normal colon tissue, and malignant colon tissues of (b) well-, (c) moderately-, and (d) poorly-differentiable colon cancer grades as observed under microscope.

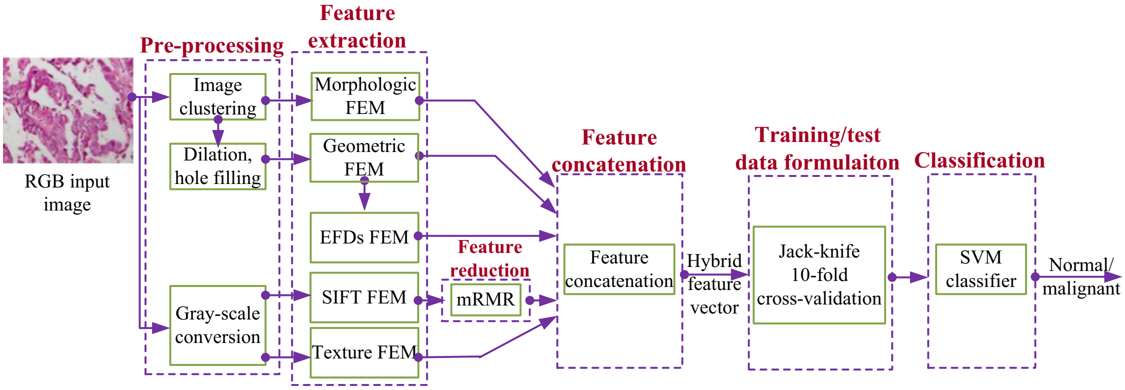


Fig. 2. Schematic diagram of the proposed HFS-CC scheme.

training/testing data formulation, and classification of images into normal or malignant categories by using SVM classifier. Fig. 2 presents top-level architecture of the proposed scheme.

In the proposed HFS-CC scheme, the hybrid features are extracted from the given dataset after a few pre-processing steps. The features are then combined for subsequent use in SVM based classification.

3.1. Pre processing

The main purpose of this stage is to make the dataset suitable for subsequent operations. Two types of preprocessing are applied on input images depending upon the requirements of subsequent feature extraction techniques. These preprocessing methods are explained in the following text.

3.1.1. Pre-processing for morphological and geometric features

Morphological and geometric features are extracted from image clusters. This clustering is achieved by running K-Means algorithm. K-Means is a non-parametric statistically iterative method, originally developed by *Fukunaga* et al. for estimation of gradients of a density function [19], and has extensive use in computer vision for image clustering [20,21]. In this study, K-Means algorithm has been applied on color intensities of pixels, and image pixels have been segregated into three clusters. The clusters depict pink region (connective tissue components), white region (luminal structures and epithelial cells), and purple region (non-epithelial cells) in the image. The clusters are then transformed to binary format using global thresholding method. Morphological features have been computed from binary clusters, whereas the binary clusters have been further dilated using Eq. (1) for geometric features.

$$f = f \oplus b \tag{1}$$

where f is the cluster and b is the square structuring element of size ‘2’. Fig. 3 shows a colon biopsy image and its corresponding white, pink, and purple clusters achieved after preprocessing.

K-Means divide an image into clusters, but does not separately identifie pink, white and purple clusters. Therefore, average intensity

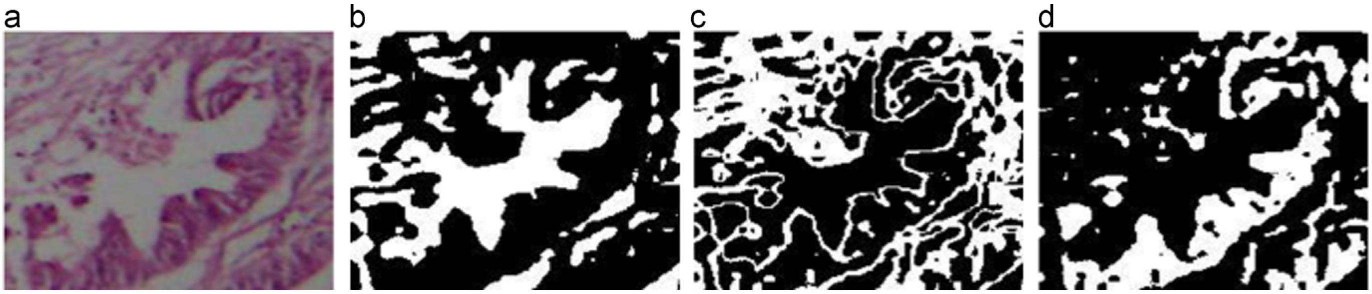


Fig. 3. Image clustering: (a) colon biopsy image, (b) white cluster, (c) pink cluster, and (d) purple cluster.

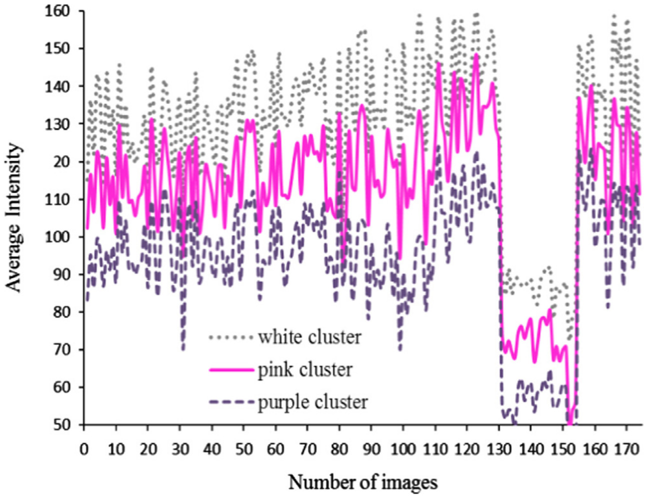


Fig. 4. Average gray-level values for white, pink and purple clusters.

values of the clusters are calculated for an image in order to identify the clusters. A white cluster has highest average intensity, a pink cluster has moderate value, whereas a purple cluster has lowest value of average intensity. So, the clusters of an image can be easily distinguished based on this parameter. Fig. 4 demonstrates mean gray level values for the three clusters.

3.1.2. *Pre-processing for SIFT and texture features*

Second type of pre-processing is specific to the extraction of SIFT and texture features. For these features, image is converted into grayscale in the preprocessing phase.

3.2. *Feature extraction*

Features provide a mean to translate an image pattern into a set of discriminatory quantitative values. The ultimate aim of this stage is to formulate a feature vector for every colon biopsy image. The individual features are combined to form a composite feature vector, which is used for image classification. Feature extraction process has already been explained in Fig. 2. There are five feature extraction modules (FEM) corresponding to individual feature extraction strategies. These modules independently extract diverse features from an image with one exception of elliptic Fourier descriptors based FEM that utilizes some information supplied by geometric FEM.

3.2.1. *Morphological features*

Morphology of tissues plays a pivotal role in determining whether tissues are normal or malignant. Morphological features provide a way to convert image morphology into a set of quantitative values. They have broadly been used in classification [22–24], segmentation [25], and so on.

Morphological FEM takes binary image clusters as input, and finds connected components in the clusters. It retains those connected components having area greater than a certain threshold T (see Section 5.3.2). The number of connected components satisfying threshold criterion may vary for different clusters, and are represented by C_w, C_p and C_r for white, pink and purple clusters, respectively. Nine morphological features, namely, area (a), perimeter (p), eccentricity (y), Euler number (l), convex area (x), compactness (o), orientation (e), length of major (m_1) and minor axes (m_2) are computed for each connected component of a cluster. Table 2 describes these features; definitions of morphological features have been taken from *Gonzalez* [26].

Eq. (2) presents morphological feature vectors for individual connected components.

$$\begin{aligned} q_i^w &= [a\ p\ y\ l\ m_1\ m_2\ x\ o\ e]_i^T \quad \text{where } i = 1, 2, 3, \ldots, Cw \\ q_i^p &= [a\ p\ y\ l\ m_1\ m_2\ x\ o\ e]_i^T \quad \text{where } i = 1, 2, 3, \ldots, Cp \\ q_i^r &= [a\ p\ y\ l\ m_1\ m_2\ x\ o\ e]_i^T \quad \text{where } i = 1, 2, 3, \ldots, Cr \end{aligned} \tag{2}$$

where $\boldsymbol{q_i^w}$, $\boldsymbol{q_i^p}$ and $\boldsymbol{q_i^r}$ are morphological feature vectors for ith connected component of white, pink and purple clusters, respectively.

Average values of the nine features obtained from connected components of a cluster constitute morphological feature vector for the corresponding cluster. The feature vectors $\boldsymbol{w}$, $\boldsymbol{p}$, $\boldsymbol{r}$ for the white, pink, and purple clusters are given in the following equation:

$$\begin{aligned} \boldsymbol{w} &= \tfrac{1}{C_w}\left[\sum_{i=1}^{C_w} q_i^w\right] \\ \boldsymbol{p} &= \tfrac{1}{C_p}\left[\sum_{i=1}^{C_p} q_i^p\right] \\ \boldsymbol{r} &= \tfrac{1}{C_r}\left[\sum_{i=1}^{C_r} q_i^r\right] \end{aligned} \tag{3}$$

The vectors $\boldsymbol{w}$, $\boldsymbol{p}$ and $\boldsymbol{r}$ are combined to form morphological feature vector $\boldsymbol{m}$.

$$\boldsymbol{m} = [\boldsymbol{w}^T \boldsymbol{p}^T \boldsymbol{r}^T]^T \tag{4}$$

3.2.2. *Texture features*

Texture features have been quite successfully used in solving classification related problems [27–29], and especially the classification of colon biopsies [11,12]. In this work, texture features have been calculated from the GLCM matrix, which encapsulates the spatial relationship between pixels of an image. Each entry (i, j)th in the GLCM defines how many times the pixel with intensity value i co-occur in a specified relationship with pixel having intensity value j. The relationship is in terms of two parameters, which are the relative distance (d) between the pixel of interest and the neighboring pixel, and their relative orientation θ. Normally, θ is quantized in four directions (0°, 45°, 90°, 135°) [10]. In

Please cite this article as: S. Rathore, et al., Automated colon cancer detection using hybrid of novel geometric features and some traditional features, Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.03.004

Table 2
Morphological features.

Feature	Definition
Area (a)	The number of pixels in a region.
Compactness (o)	It is defined as (perimeter)^2/area.
Convex area (x)	The number of pixels in convex image.
Eccentricity (y)	The ratio of the distance between the foci of the ellipse and its major axis length.
Euler number (l)	The number of objects in the region minus number of holes in those objects.
Major axis (m_1)	The length of the major axis of the ellipse. Measured in number of pixels.
Minor axis (m_2)	The length of the minor axis of the ellipse. Measured in number of pixels.
Orientation (e)	The angle between the horizontal axis and the object.
Perimeter (p)	The number of pixels in the boundary of the region.

Table 3
Texture features.

Feature	Equation	Definition
Contrast	$t=\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 p_{ij}$	Measures the intensity contrast between a pixel and its neighbor over the entire image.
Correlation	$\rho=\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(i-m_i)(j-m_j)p_{ij}}{\sigma_i\sigma_j}$	Measures the degree of correlation between a pixel and its neighbors over the entire image.
Energy	$n=\sum_{i=1}^{K}\sum_{j=1}^{K}{p_{ij}}^2$	Measures uniformity in the image.
Homogeneity	$h=\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{p_{ij}}{1+\lvert i-j\rvert}$	Measures the spatial closeness of the distribution of elements in **G** to the diagonal.
Randomness	$r=-\sum_{i=1}^{K}\sum_{j=1}^{K}p_{ij}\log_2 p_{ij}$	Measures the randomness of the elements of gray-level co-occurrence matrix.

this work, texture features have been computed at four possible directions, and two distances ($d=1,2$), thereby resulting in a total of 8 GLCM matrices. The values of the GLCM matrices have been averaged in four directions for each distance d, thereby resulting in two GLCM matrices (for $d=1$ and $d=2$). The texture features of randomness (r), contrast (t), correlation (ρ), energy (n), and homogeneity (h) have been computed from the two GLCM matrices. The results in the experimental section correspond to the GLCM that leads to maximum classification capability, which is the four-directional average GLCM computed at $d=1$.

The initial eight GLCM matrices have also been used separately for classification of colon image dataset. Furthermore, several combinations of these GLCMs, and a hybrid of the features computed from these GLCMs have also been investigated. But, the best results have been achieved for the average four-directional GLCM computed at $d=1$. Table 3 formulates definitions of texture features and mathematical formulae for their calculation, as given in *Gonzalez* and *Woods* [26].

These features are computed from gray-level co-occurrence matrix **G**, where i and j represent indices of its rows and columns. p_{ij} is the ijth term of **G** divided by the sum of its elements. The terms m_i and m_j are the mean, σ_i and σ_j are the standard deviation of ith row and jth column of **G**. Above mentioned features are combined to form texture feature vector $\boldsymbol{t}$.

$$\boldsymbol{t}=[r\ t\ \rho\ n\ h]^T \tag{5}$$

3.2.3. SIFT features

SIFT features, originally proposed by Lowe [30], are normally used in problems of panoramas reconstruction [31], face identification/authentication [32–34], and most importantly, visual object tracking [35]. However, its diverse and distinctive nature helps researchers to use it in other application areas. SIFT features are robust against image scaling, rotation, illumination changes, noise, and blurry effects. These properties of SIFT features make them a good choice for classification of colon samples.

SIFT features are extracted through a staged filtering process. In the first step, key points are localized in an image. To this end, original image is repeatedly convolved with Gaussian to produce convolved scale-space images. Then, adjacent Gaussian convolved images are subtracted to produce difference of Gaussian images. After calculating difference at one scale, original image is down-sampled by a factor of 2, and process is repeated until reaching lowest possible scale. First step detects large number of key points, which are reduced in the next step. In the second step, each pixel is matched against 8 neighbors in its own scale and 9 neighbors in scales above and below it. The points having their value either greater or smaller compared to all the neighboring pixels retain after this step. In the third step, unstable key points i.e. the points that are poorly localized along edges or have poor contrast are discarded.

Finally, orientations and descriptors are assigned to the remaining key points. Orientations are assigned based on directions of local image gradient. For calculation of descriptors, magnitude and orientation values of pixels lying in 16 × 16 window around a given pixel are used to compute 16 orientation histograms. These histograms contain samples from 4 × 4 sub-regions of the given window, and have 8 bins each. The magnitudes and orientations are further smoothed by a Gaussian function with equal to one half the width of the descriptor window. The descriptor then becomes a vector of all the values of these histograms. Since there are 16 histograms with 8 bins each, therefore, the vector has 128 elements. Descriptor vector is finally normalized to make it independent of linear and non-linear illumination changes.

In this particular research study, SIFT features are calculated for the given dataset. SIFT feature vector for an image comprises orientations and descriptors of SIFT key points of that particular image. The number of SIFT key points may vary from image to image, therefore, most distinctive S key points are picked from each image to make equal sized feature vectors. The libsiftfast-1.2 library [36] has been used for feature extraction. SIFT feature

Please cite this article as: S. Rathore, et al., Automated colon cancer detection using hybrid of novel geometric features and some traditional features, Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.03.004

vector has the following composition.

$$\begin{aligned} &\boldsymbol{\theta} = [\theta_1\ \theta_2\ \theta_3 \ldots \theta_S] \\ &\boldsymbol{q}_i = [q_1\ q_2\ q_3\ \ldots\ q_{128}]_i^T \text{ where } i = 1, 2, 3, \ldots, S \\ &\boldsymbol{s}^* = [\boldsymbol{\theta}\ \boldsymbol{q}_1^T\ \boldsymbol{q}_2^T\ \boldsymbol{q}_3^T \ldots \boldsymbol{q}_S^T]^T \end{aligned} \tag{6}$$

In Eq. (6), $\boldsymbol{\theta}$ contains orientations of S points, whereas $\boldsymbol{q}_i$ vector contains features of *i*th SIFT point. Finally orientations and features of S points are combined to form SIFT fetaure vector $\boldsymbol{s}^*$.

3.2.3.1. SIFT feature reduction. In classification problems, huge size, imbalanced nature and high dimensionality of training dataset mainly cause the classification algorithms to suffer in accurately predicting the samples. Therefore, data must be reduced in size prior training a classifier. In our classification problem, SIFT features have quite large size. Therefore, in order to alleviate the curse of dimensionality issues, SIFT features have been reduced by using minimum redundancy and maximum relevance (mRMR) method [37]. Fig. 5 demonstrates dimensionality reduction process for SIFT features. mRMR model takes original SIFT feature vector $\boldsymbol{s}^*$ and corresponding image labels $b_1, b_2, \ldots, b_{174}$ as input, and returns reduced feature vector $\boldsymbol{s}$.

3.2.4. *Proposed geometric features*

In a normal colon tissue, epithelial cells and lumen are nearly elliptic, whereas, in a malignant colon tissue, epithelial cells and lumen merge together, thereby resulting in irregular shaped white regions. Furthermore, the epithelial cells in a normal colon tissue are elliptic, and have symmetry in their sizes and distribution. On the other hand, the epithelial cells in a malignant colon tissue are irregular, and have no symmetry in their distribution and sizes. Normal and malignant colon tissues are shown in Fig. 6 for comparison.

In this research study, we speculated that this variation can be exploited to identify normal and malignant colon tissues. Therefore, geometric features have been proposed with an intention to capture geometrical differences between the structure of epithelial cells and lumen in normal and malignant colon tissues.

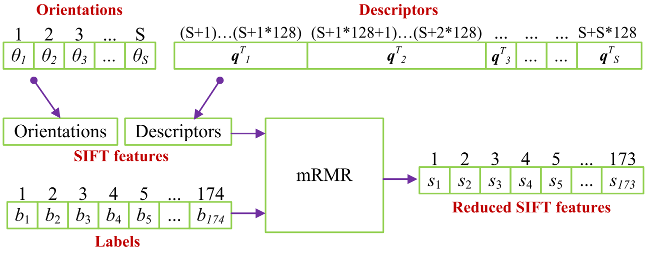


Fig. 5. The process of selecting discerning features from SIFT features.

The calculation of geometric features is a three step process. In the first step, elliptic objects (epithelial cells and lumen) are detected in colon biopsy images. In the second step, the detected objects are divided into different categories. In the third step, the information about the size and spatial distribution of these objects is exploited to calculate geometric features. These steps are explained in the following text.

3.2.4.1. Object detection. The first step is to locate elliptic objects in colon biopsy images. Since epithelial cells and lumen belong to white cluster, therefore, elliptic objects are detected only in the white cluster of colon biopsy images. A novel algorithm has been proposed for this purpose, which is shown in Fig. 7.

In the proposed algorithm, connected components are generated in white cluster of colon biopsy images. Smaller components, which have arisen due to blur in colon biopsy images and have area smaller than component area threshold (CAT), are excluded from further experimentation. Later, each connected component is processed individually, and elliptic objects are found in the component.

The process of searching elliptic objects in a single component is shown in the right half of Fig. 7. Initially, four patterns of ellipses i.e. horizontal (0°), vertical (90°), diagonal (45°), and off-diagonal (135°) as shown in Fig. 8 are generated starting with maximum values of semi-major (SMJA) and semi-minor axes (SMIA).

Generated ellipses are in fact matrices, and their size depends upon length of semi-major and semi-minor axes. For example, for 5 units' long semi-major axis and 3 units' long semi-minor axis, horizontal, vertical, diagonal and off-diagonal ellipses are matrices of size 11 × 7, 7 × 11, 11 × 11, and 11 × 11, respectively. The pixels lying inside ellipse have value '1', whereas outer pixels have value '0' as demonstrated in Fig. 8.

The four generated ellipses are found in each connected component of a cluster. For the reason, system traverses pixels of the connected component one at a time, and extracts four regions

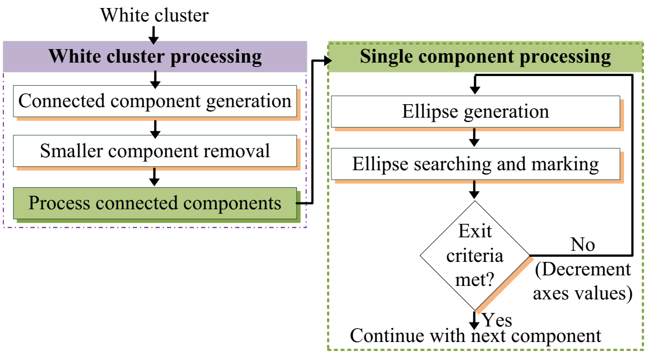

White cluster
White cluster processing
Connected component generation
Smaller component removal
Process connected components
Single component processing
Ellipse generation
Ellipse searching and marking
Exit criteria met?
No
(Decrement axes values)
Yes
Continue with next component

Fig. 7. The process of detecting elliptic shape based epithelial cells in white cluster of colon biopsy images.

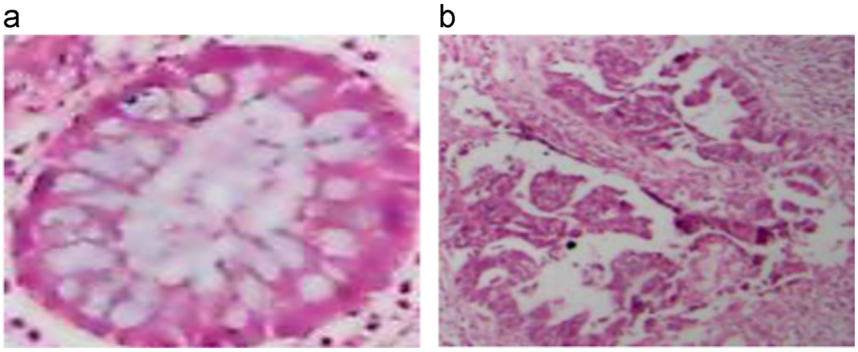
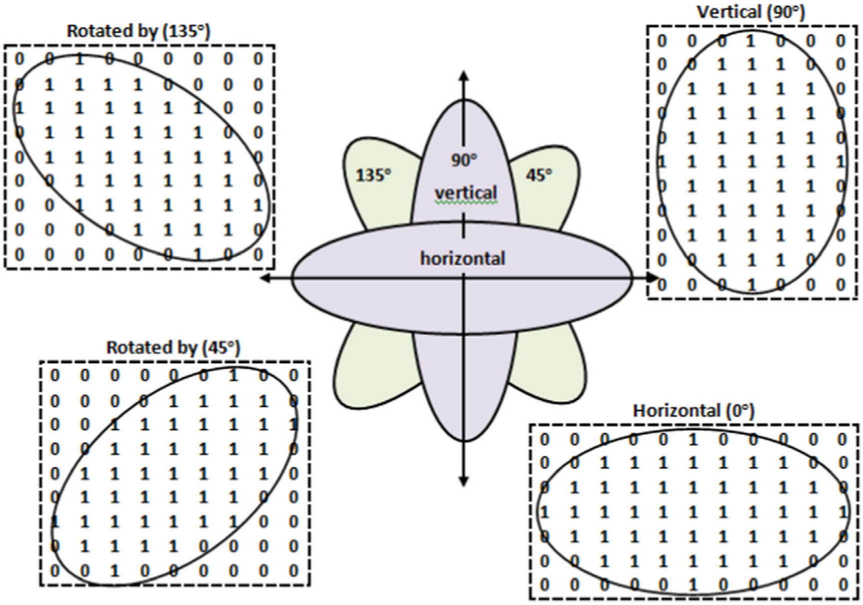


Fig. 6. Histopathological images of (a) normal and (b) malignant colon tissue.



Rotated by (135°)

0	0	1	0	0	0	0	0	0
0	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	0	0
0	1	1	1	1	1	1	0	0
0	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	1
0	0	0	0	1	1	1	1	0
0	0	0	0	0	0	1	0	0

Vertical (90°)

0	0	0	1	0	0	0
0	0	1	1	1	0	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
1	1	1	1	1	1	1
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	1	1	1	1	1	0
0	0	1	1	1	0	0
0	0	0	1	0	0	0

Rotated by (45°)

0	0	0	0	0	0	1	0	0
0	0	0	0	1	1	1	1	0
0	0	1	1	1	1	1	1	1
0	0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0	0
1	1	1	1	1	1	1	0	0
0	1	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0

Horizontal (0°)

0	0	0	0	0	1	0	0	0	0	0
0	0	1	1	1	1	1	1	1	0	0
0	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	1	0	0
0	0	0	0	0	1	0	0	0	0	0

Fig. 8. Horizontal, vertical, and diagonal ellipses (semi-major axis=5, semi-minor axis=3).

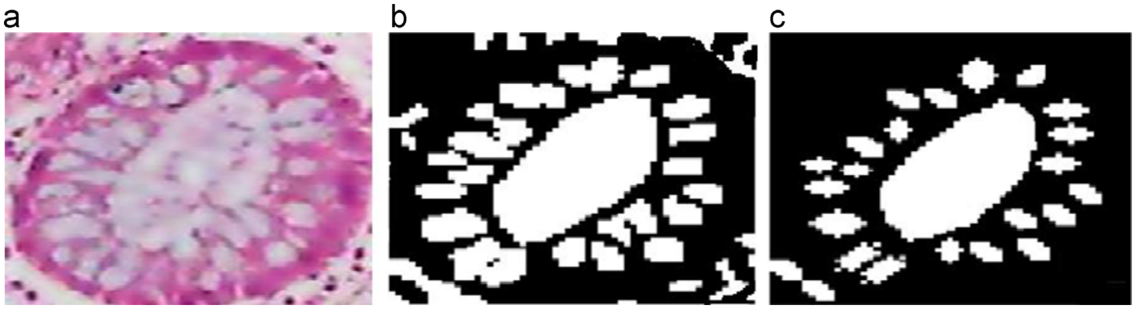


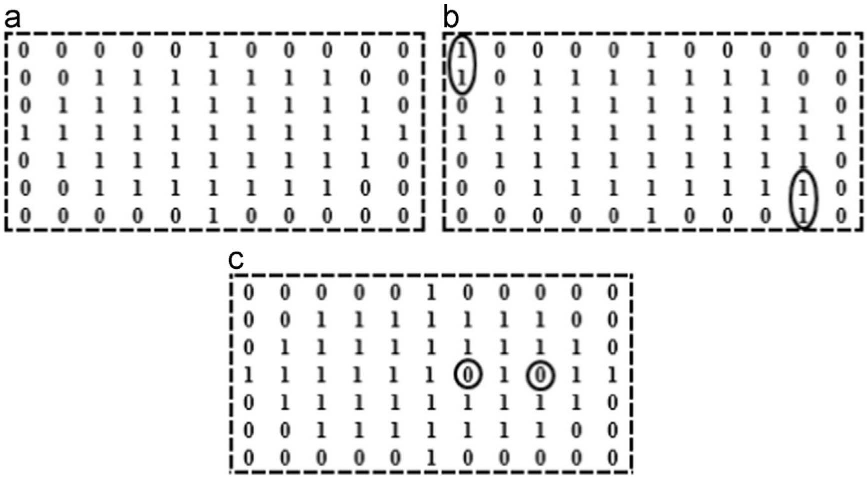
Fig. 9. Ellipses detected in white cluster of a colon biopsy image: (a) colon biopsy image, (b) white cluster, (c) detected elliptic objects.

of the same size as horizontal, vertical, diagonal and off-diagonal ellipses around the pixel. The pixel values of horizontal, vertical, diagonal, and off-diagonal windows are compared with respective pixel values of the horizontal (0°), vertical (90°), diagonal (45°), and off-diagonal (135°) ellipses. If all the pixels having value '1' in an ellipse are also '1' in the extracted window, an ellipse of that particular orientation is supposed to exist at the pixel. Once the system finishes traversing all the pixels of a connected component, the object detection process is repeated with the remaining pixels (which have not been assigned to any elliptic object) of the component by decrementing the values of SMJA and SMIA by one unit. The process is continued for the same connected component provided there are some unassigned pixels, and the SMJA and SMIA have not reached minimum bounds. Otherwise, the process is started for the next connected component of the cluster in the same fashion. Fig. 9 demonstrates the elliptic objects detected in the white cluster of a colon biopsy image.

The blur or noise in colon biopsy images may disturb the functioning of K-Means, and some inner pixels of epithelial cells may be 0, thereby leading to not exact but almost elliptic shapes in the clusters. Furthermore, the ellipse searching process in four orientations covers most of the ellipses, but there may be some ellipses which are slightly tilted from the four defined orientations. Therefore, a concept of membership function has been introduced in the proposed HFS-CC technique in order to find nearly elliptic shape based epithelial cells and the epithelial cells tilted from four standard orientations. The membership function defines the percentage of pixels, which have the same value in generated ellipse and the extracted window. The optimal value of membership, found through experimentation, is 95%. Fig. 10 demonstrates membership function. In Fig. 10(a), there is a horizontal ellipse that needs to be found in the image patches (b) and (c). A full match has been found in the pattern given in Fig. 10(b). Circled numbers in Fig. 10(b) represent those image pixels which have value 1, but they do not participate in matching process because they lie outside the boundary of generated ellipse. A partial match has been found in Fig. 10(c). It shows nearly elliptic shape where two circled numbers inside ellipse are 0, however membership function helps detecting such ellipses.

3.2.4.2. Object categorization. The detected objects are further divided into two categories depending upon object area threshold (OAT). The largest object, which is lumen in case of normal images, and is either lumen or a larger part of scattered lumen in case of malignant images, is not divided into any of the categories. The objects having size larger than OAT belong to one category (object type-1), whereas, objects having size smaller than OAT belong to second category (object type-2). Fig. 11(a) shows the categorization of objects into two object types. The objects having size greater than OAT are in yellow, and the objects having size smaller than OAT are in blue.

3.2.4.3. Calculation of geometric features. Geometric features are computed based on the information about size and spatial distribution of the detected objects. These features include:



a

0	0	0	0	0	1	0	0	0	0	0
0	0	1	1	1	1	1	1	1	0	0
0	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	1	0	0
0	0	0	0	0	1	0	0	0	0	0

b

1	0	0	0	0	1	0	0	0	0	0
1	0	1	1	1	1	1	1	1	0	0
0	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	1	1	0
0	0	0	0	0	1	0	0	0	1	0

c

0	0	0	0	0	1	0	0	0	0	0
0	0	1	1	1	1	1	1	1	0	0
0	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	0	1	0	1	1
0	1	1	1	1	1	1	1	1	1	0
0	0	1	1	1	1	1	1	1	0	0
0	0	0	0	0	1	0	0	0	0	0

Fig. 10. (a) Simulated horizontal ellipse, (b) image pattern (full pattern match), and (c) image pattern (partial match 95.6%).

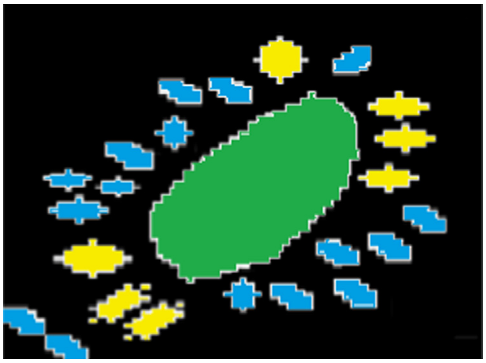


Fig. 11. Object categorization into two object types based on OAT.

1. **Object size uniformity (OSU):** OSU measures the uniformity in the sizes of objects both at the local and global level. The local features are called local object size uniformity (LOSU), and global features are called global object size uniformity (GOSU).
   Local object size uniformity: In order to calculate LOSU features, a circular window is iterated on each image pixel, and standard deviation of areas of both the object types lying within the window is separately calculated. For instance, in window 2 in Fig. 12(a), there are 2 objects of object type 1, and 4 objects of object type 2. The LOSU features are computed by calculating the standard deviation of areas for the 2 and 4 objects, respectively, for object type 1 and 2. Since both the object types have areas of different scales, therefore, LOSU features are normalized by dividing standard deviation of areas with respective mean for each particular object type. Two features per pixel are computed this way. The LOSU features will be zero for pixels, which have either no object or only one object in a circular window around them, therefore, these pixels do not contribute in the calculation of geometric features. This scenario is shown in window 1 of Fig. 12(a). In the end, we do average the non-zero values of features to compute two LOSU features per image.
   Since the detected objects have almost the same area in normal colon biopsy images, therefore, the value of LOSU features will be near to zero for these images. On the other hand, since the detected objects greatly differ in terms of area in malignant colon biopsy images, therefore, LOSU features will have larger values for these images. These features are calculated using smaller and larger window of radii RS and RL, respectively, and are named as $LOSU_1$ and $LOSU_2$ for smaller window, and $LOSU_3$ and $LOSU_4$ for larger window, respectively.
   Global object size uniformity: In order to calculate the GOSU features, standard deviation of the area of all the detected objects for a particular object type is calculated. This measure will also be small for normal colon biopsy images, and vice versa due to more uniformity in the detected objects of normal colon biopsy image. These features are named as $GOSU_1$ and $GOSU_2$.
2. **Object spatial distribution uniformity (OSDU):** OSDU feature measures the uniformity in the spatial distribution of detected objects. It is a measure of the magnitude of sum of position vectors for all the objects other than lumen. The position vectors are calculated with reference to the centre of the lumen (largest detected object). The process of computing OSDU features from a colon biopsy image is shown in Fig. 12(b). The value of OSDU feature will be near to zero for normal colon biopsy images since the detected objects are uniformly distributed in space around the lumen. On the other hand, OSDU feature will have larger values for malignant colon biopsy images since the detected objects do not follow any standard spatial distribution. The OSDU feature is not calculated at the local level since the objects do not have any regular orientation at the local level.

Eq. (7) depicts composition of geometric feature vector $\mathbf{g}$.

$$\mathbf{g} = [LOSU_1 \;\; LOSU_2 \;\; LOSU_3 \;\; LOSU_4 \;\; GOSU_1 \;\; GOSU_2 \;\; OSDU] \tag{7}$$

The optimal values of several parameter such as RS, RL, membership function, maximum and minimum bounds for SMJA and SMIA, OAT, and CAT are calculated. The process of calculating these values is given in explained in Section 5.3.2.

3.2.5. Elliptic Fourier descriptors

Lumen and epithelial cells have elliptic shape, therefore, it is speculated that elliptic Fourier descriptors (EFDs) of these constituents will help in discriminating normal and malignant colon tissues. EFDs were initially introduced in 1982 by Kuhl et al. for classification of several solid objects such as windmill, tank, etc. [38]. However, later on EFDs have been extensively used in pattern recognition [39,40].

The computation of EFD features is a two step process. In the first step, elliptic objects are detected in the white cluster of colon biopsy images as already discussed in Section 3.2.4.1. In the second step, elliptic objects are sorted in descending order based on their

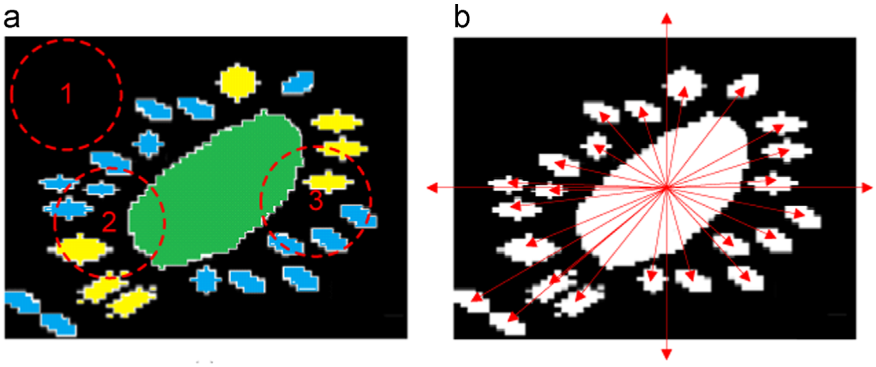


Fig. 12. The process of computing geometric features: (a) OSU features and (b) OSDU features.

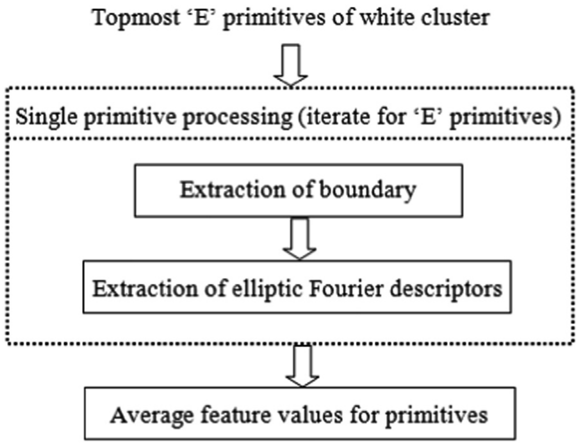

Topmost 'E' primitives of white cluster
Single primitive processing (iterate for 'E' primitives)
Extraction of boundary
Extraction of elliptic Fourier descriptors
Average feature values for primitives

Fig. 13. EFDs extraction.

area, and EFDs of the top-most E objects are calculated up to the desired harmonic level H. Selection of optimal values for E and H is explained in Section 5.3.2. The process of extraction of EFDs is elaborated in Fig. 13.

EFDs are based on chain codes, which approximate the shape of a closed contour by a sequence of eight standardized line segments, and are invariant to dilation, translation, rotation and starting point of a contour. EFDs of a closed contour comprise x and y-projection of its chain codes. H harmonic levels are used for extraction of EFDs, and there are four Fourier coefficients i.e. *a, b, c* and *d* against each harmonic level. Eq. (8) presents elliptic Fourier descriptors for E elliptic primitives.

$$\begin{aligned} \boldsymbol{a}_i &= [a_1\ a_2\ a_3 \ldots a_H]_i^T \text{ where } i = 1, 2, 3, \ldots, E \\ \boldsymbol{b}_i &= [b_1\ b_2\ b_3 \ldots b_H]_i^T \\ \boldsymbol{c}_i &= [c_1\ c_2\ c_3 \ldots c_H]_i^T \\ \boldsymbol{d}_i &= [d_1\ d_2\ d_3 \ldots\ d_H]_i^T \end{aligned} \tag{8}$$

where $\boldsymbol{a}_i$,$\boldsymbol{b}_i$, $\boldsymbol{c}_i$, and $\boldsymbol{d}_i$ vectors contain *a*, *b*, *c* and *d* Fourier coefficients of *i*th primitive upto harmonic level H. $\boldsymbol{a}_i$,$\boldsymbol{b}_i$, $\boldsymbol{c}_i$, and $\boldsymbol{d}_i$ vectors are averaged to form $\overline{\boldsymbol{a}}$, $\overline{\boldsymbol{b}}$, $\overline{\boldsymbol{c}}$and $\overline{\boldsymbol{d}}$ average vectors.

$$\overline{\boldsymbol{a}} = \frac{1}{E}\sum_{i=1}^{E}\boldsymbol{a}_i, \quad \overline{\boldsymbol{b}} = \frac{1}{E}\sum_{i=1}^{E}\boldsymbol{b}_i, \quad \overline{\boldsymbol{c}} = \frac{1}{E}\sum_{i=1}^{E}\boldsymbol{c}_i, \quad \overline{\boldsymbol{d}} = \frac{1}{E}\sum_{i=1}^{E}\boldsymbol{d}_i \tag{9}$$

Average vectors are then combined to form final elliptic feature vector ***e***.

$$\boldsymbol{e} = [\overline{\boldsymbol{a}}^T \overline{\boldsymbol{b}}^T \overline{\boldsymbol{c}}^T \overline{\boldsymbol{d}}^T]^T \tag{10}$$

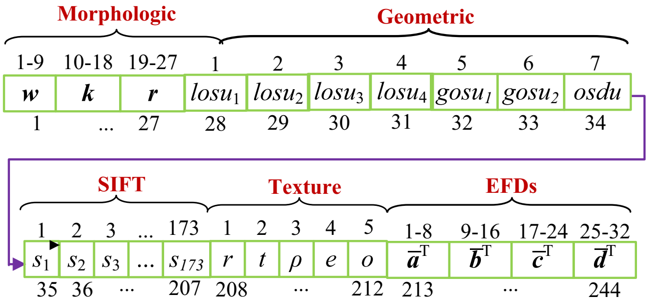


Fig. 14. Developing a hybrid feature vector by combining individual feature vectors.

3.3. Feature concatenation

Individual features are combined to make a hybrid feature vector. The features are aligned in sequence as presented in Fig. 14. Upper indices are for individual features, whereas lower indices are for hybrid features. Hybrid feature vector has 244 dimensions.

3.4. Training/testing data formulation

In this work, Jackknife 10-fold cross-validation technique has been employed for training/testing data formulation and parameter optimization. It is a commonly practiced technique that has been successfully used in the past to validate the accuracy of prediction. In 10-fold Jackknife test, data are divided into 10 folds. 9 folds participate in training, and the classes of the samples belonging to the remaining fold are predicted based on the training performed on 9 folds. The test samples in the test fold are purely unseen for the trained model. This sampling process is repeated 10 times and the class of each sample is predicted. Finally, the predicted labels of the unseen test samples are used to determine classification accuracy. The Jack-knife process is repeated for each combination of system's parameters, and classification performance has been reported for the sample that leads to maximum classification accuracy on the unseen test data. Fig. 15 presents Jackknife 10-fold cross-validation process for the calculation of classification performance of morphological feature vector using linear SVM. There are two parameters involved in this task; one is the area threshold of morphological features (T), and second is the constraint violation cost (*c*) of linear SVM. The parameters have been varied in their potential ranges, and Jack-knife process is repeated for each combination. The classification accuracy on unseen test data is measured for each combination of parameter values, and the best achieved classification accuracy has been reported in Section 5.

3.5. Classification

In recent times, SVM classifier has received noticeable attention for attaining higher classification success. SVM classifier is considered a better performer compared to other classifiers. It has been quite successfully used in different application areas of medical diagnosis [41–43]. In the training phase of SVM, it maps non-separable data to a high dimensional space where it becomes linearly separable. In high dimensional space, SVM creates a partition surface between data of two classes while trying to maximize the margin of separation between classes. Decision surface divides total feature space into two sub-spaces where each sub-space belongs to single class. In the testing phase, test data is mapped to the space and labels are assigned to the images depending upon the sub-space in which their features lie. Linear, RBF and Sigmoid kernels have been used for classification, and classification accuracy is evaluated using 10-fold cross-validation.

4. Performance evaluation measures

In our experiments, we provide visual results obtained by the algorithms i.e. labels assigned to the samples during testing. Further, we quantitatively evaluate the results using well-known performance metrics such as accuracy, sensitivity, specificity, Matthews's correlation coefficient (MCC), F-score and receiver operating characteristics (ROC) curves. The calculation of parameters involves true positive (TP), false positive (FP), true negative (TN), and false negative (FN). True negative and true positive are the number of correctly classified negative and positive samples, whereas, false negative and false positive are the number of positive and negative samples, which are incorrectly classified.

4.1. Accuracy

Accuracy is a measure of overall effectiveness/usefulness of the classification scheme. It can be calculated using equation given below.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100 \tag{11}$$

4.2. Sensitivity

Sensitivity is used to measure the ability of a classifier to recognize patterns of positive class. It can be obtained using the following equation.

$$Sensitivity = \frac{TP}{TP+FN} \tag{12}$$

4.3. Specificity

Specificity is calculated to measure the ability of a classifier to recognize patterns of negative class. The following equation is used to calculate specificity.

$$Specificity = \frac{TN}{TN+FP} \tag{13}$$

4.4. Matthews correlation coefficient

MCC serves as a measure of classification in binary class problems. Its value ranges from −1 to +1. +1 means classifier always predicts a right label, whereas −1 means classifier always commits a mistake. However, 0 means random prediction. MCC

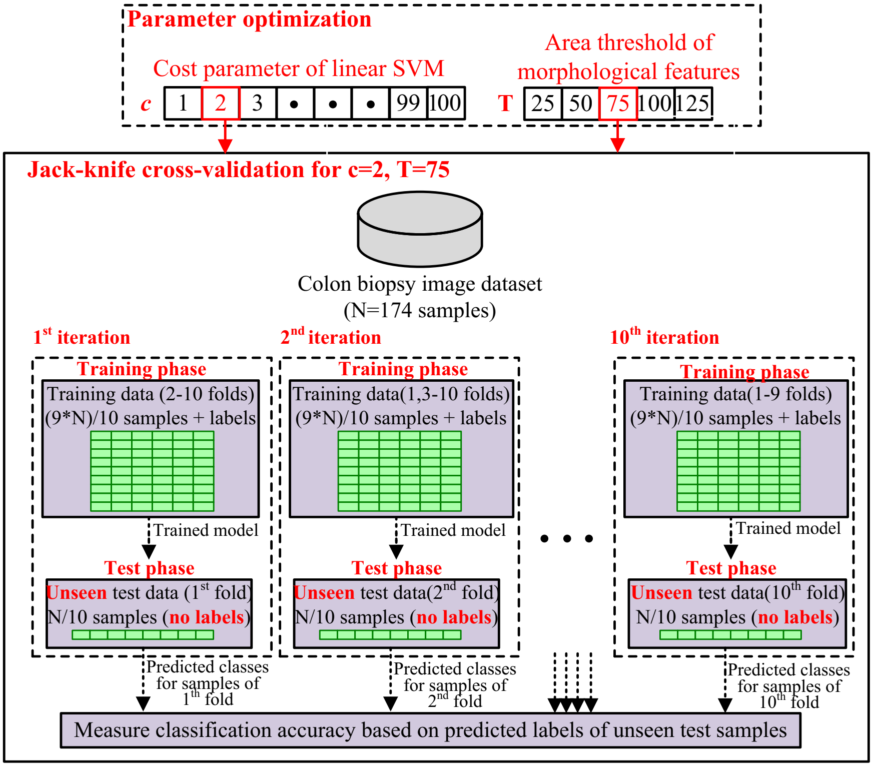

Parameter optimization
Cost parameter of linear SVM
c
1
2
3
•
•
•
99
100
Area threshold of morphological features
T
25
50
75
100
125
Jack-knife cross-validation for c=2, T=75
Colon biopsy image dataset (N=174 samples)
1st iteration
2nd iteration
10th iteration
Training phase
Training data (2-10 folds) (9*N)/10 samples + labels
Training data(1,3-10 folds) (9*N)/10 samples + labels
Training data(1-9 folds) (9*N)/10 samples + labels
Trained model
Test phase
Unseen test data (1st fold) N/10 samples (no labels)
Unseen test data(2nd fold) N/10 samples (no labels)
Unseen test data(10th fold) N/10 samples (no labels)
Predicted classes for samples of 1th fold
Predicted classes for samples of 2nd fold
Predicted classes for samples of 10th fold
Measure classification accuracy based on predicted labels of unseen test samples

Fig. 15. Jack-knife cross-validation for determining classification capability of morphological features for $c=2$ and $T=75$. The figure shows Jack-knife process for linear SVM.

can be calculated using the following formula.

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{((TP+FN)(TP+FP)(TN+FN)(TN+FP))}} \tag{14}$$

4.5. F-score

F-score makes use of precision and recall to calculate accuracy of classification.

$$Precision=\frac{TP}{TP+FP}, \quad Recall=\frac{TP}{TP+FN}$$

The F-score can be calculated by using Eq. (15). It is weighted average of precision and recall values. Its value ranges between 0 and 1, where 0 is the worst possible score and 1 is the best possible.

$$Fscore=2\times\frac{Precision\times Recall}{Precision+Recall} \tag{15}$$

4.6. ROC curves

An ROC curve is a standard way for graphical representation of the classification performance of a system [44]. It characterizes the system over its entire operating range, and is created by plotting true positive rate (TPR) against false positive rate (FPR). TPR represents the number of correct positive cases divided by the total number of positive cases. FPR, on the other hand, is the number of negative cases predicted as positive cases divided by the total number of negative cases.

5. Results and discussions

The proposed technique has been tested to classify colon biopsy images into normal and malignant samples. The proposed and traditional features are extracted from colon biopsy images, and are combined to be used in the classification process. All of the computations have been performed on Intel Core i7 PC.

5.1. Dataset

Our dataset comprises 68 colon biopsy samples taken from randomly selected patients of Rawalpindi Medical College (RMC), Pakistan. Samples comprise 5–6 μm thick tissue section, stained with H&E (Hematoxylin & Eosin). Imaging equipment was provided by PAEC General Hospital, Islamabad. The resolution and magnification factors of microscope were set to 600 × 800 and 10 ×, respectively. A dataset of 174 RGB images has been acquired from the samples. In order to eliminate the risk of inter-observer variability in the diagnosis of colon biopsy slides, ground truth labels have been assigned to the images by four classified histopathologists, namely, Dr. Imtiaz Qureshi (RMC), Dr. Rahat Abbas (RMC), Brig. Shoaib Nayyar Hashmi (Armed Forces Institute of Pathology, Rawalpindi), Prof. Dr. Anwar-ul-Haq (Azad Jammu and Kashmir Medical College, Muzaffarabad). Among the 174 images, the pathologists have perfect agreement for 170 images. For the controversial four images, there exists 75% agreement i.e. three out of the four pathologists have assigned the same labels. The final labels have been assigned to the images based on majority voting. The Kappa statistic has also been calculated in order to determine variability in the diagnosis of different pathologists. The Kappa value for the overall dataset is 0.9768, which shows good agreement between the diagnosis of different pathologists in case of this dataset. The confidentiality of the patients has been sustained right through this research work. The college has provided the details about gender and age of the patients only. The information about the age and gender of patients, and categorization of images into multiple classes is summarized in Table 4.

5.2. Experimental setup

Experimentation starts with the selection of appropriate values of parameters for different feature extraction strategies. The selection of optimal values has been discussed in Section 5.3. The optimal values obtained this way are used in subsequent sections. Sections 5.4 and 5.5, respectively, analyze the performance of proposed individual and hybrid features. Three variants of SVM classifier (linear, RBF, Sigmoid) have been employed to classify given feature set into normal and malignant classes. Jackknife 10-fold cross validation has been used, and data has been scaled in the range 0–1 prior to classification. The classification accuracies as already described in Section 3.4 have been determined based on the labels of the unseen test samples (test folds) in different iterations. Section 5.6 describes the CPU time involved in feature extraction and classification of different feature extraction strategies.

The experiments reported in Sections 5.3–5.6 have been performed by applying 10-fold Jack-Knife on complete dataset. Later, the dataset has been divided into separate training and test data, and the same experiments have been performed. In this context, the performance of individual and hybrid feature sets has been summarized in Section 5.7. Section 5.8 provides a performance comparison of the proposed technique with existing colon biopsy image based classification techniques.

5.3. Selection of optimal values for system parameters

The performance of the proposed HFS-CC system depends on several parameters, which need to be tuned for optimal performance. In the subsequent text, analysis of optimal values of SVM models and feature selection methods has been presented in detail. Only RBF SVM has been used, and the parameter values have been obtained by taking classification accuracy into account. These optimal values have been used in the experimental results shown in Sections 5.4–5.6.

Table 4
Statistics for the dataset.

Parameters	Values
Number of images	174
Distribution of images	92 malignant, 82 normal
Grades of malignant images	23 poor-, 44 moderate-, and 25 well-differentiable
Age of patients	42–68, Mean = 57.11, Standard deviation = 6.35
Age of female patients	43–63
Age of male patients	42–68
Kappa statistic for inter-observer variability	0.9768

5.3.1. SVM models

The performance of SVM classifiers depends on several parameters. In this research study, grid search method [45] has been employed for selection of optimal parameter values by carefully setting grid range and step size. Linear kernel involves only one parameter ('*c*' soft margin constant), which is the constraint violation cost associated with the data point occurring on the wrong side of the decision surface. The parameter γ is involved in RBF and sigmoid kernels. Its optimal value has been obtained by adjusting the grid range of $\gamma = [0.001, \ldots, 0.099]$ with $\Delta\gamma = 0.002$for both the kernels. A parameter *r* is specific to sigmoid kernel only, and its default value is used.

Similarly, there is another parameter called number of folds, which is actually the parameter of Jackknife cross-validation, but may affect the performance of SVM classifier. Therefore, its optimal value must be selected prior to classification. In this particular research study, number of folds has been varied in the range 5, 10, …,30, and the results are shown in Fig. 16 in terms of classification accuracy and computational time requirements of RBF classifier for all feature extraction techniques. Fig. 16(a) demonstrates that classification accuracy slightly varies by varying folds but the variation is negligible. On the other hand, the classification time of RBF classifier as shown in Fig. 16(b) increases with an increase in number of folds. Therefore, we have opted 10-fold cross-validation in this research work.

5.3.2. Feature extraction strategies

There are several parameters involved in feature extraction, and it is quite essential to select appropriate values for these parameters prior to classification. An experimental process has been adopted to find these values. This section explains the selection of optimal values of parameters for individual feature categories one by one.

SIFT features: The performance of SIFT features significantly depends upon number of SIFT points (*S*). Therefore, we have tried to get an insight into the relationship between S and classification accuracy. Fig. 17(a) shows the results. Initially, classification accuracy shows a transitory behavior. But, a uniform value of accuracy has been observed beyond 250 SIFT points. Therefore, 250 SIFT points have been used for classification.

Morphological features: Morphological features are extracted from connected components having size greater than an area threshold (*T*). The optimal value of *T* has been obtained by varying it from 25 to 125, and by extracting corresponding features. Classification accuracy for each feature set is given in Fig. 17(b). Classification accuracy increases with an increase in the value of *T*, but shows a stable behavior after $T=100$. Therefore, component area threshold of 100 is used in this particular research study.

Elliptic Fourier descriptors: There are two parameters involved in the extraction of EFDs: number of elliptic primitives of white cluster (*E*), and harmonics level (*H*). In order to find optimal values of these parameters, *E* and *H* have been varied in the ranges $E=5, 10, \ldots, 25$ and $H=4, 8, \ldots, 20$. Corresponding results are shown in Fig. 17(c). Results demonstrate that for fixed number of elliptic primitives, accuracy first slightly increases and then gradually decreases with increase in harmonics level. On the other hand, accuracy increases with increase in number of elliptic primitives while keeping fixed harmonic level. Therefore, one can conclude that smaller values of *H* and higher values of *E* are a better choice. Highest classification accuracy has been observed for $E=25$ and $H=8$, therefore, this combination of parameters has been used in further experimentation.

Geometric features: The performance of geometric features depends upon several parameters, namely, OAT, CAT, SMJA, SMIA, RL, RS, and membership function. As there are many parameters, therefore, it is computationally expensive to explore each combination of values of these parameters to determine an optimal combination for the given dataset. Hence, the optimal values of these parameters have been found by employing genetic algorithm (GA). GA has also been used in the past for optimization of systems' parameters [46,47]. Table 5 shows the optimal values of these parameters found through GA. The optimal values of SMJA and SMIA range from 9 to 80, and from 5 to 28, respectively. For each value of SMJA i.e. from 9 to 80, the SMIA has been varied from its lower bound up to that value of SMJA. For instance, for SMJA=10, the SMIA has been varied from 5 to 10. The ellipse detection algorithm has been iterated for each of these combinations of SMJA and SMIA.

5.4. Performance analysis of individual feature extraction strategies

In the first set of experiments, individual features have been used for classification, and performance evaluation parameters have been measured. Table 6 demonstrates corresponding results.

A reasonable performance has been observed in case of every feature extraction strategy. However, close analysis reveals that the proposed geometric features perform better in terms of most of the performance metrics regardless of the choice of SVM kernel. The proposed geometric features capture true morphology/geometry of constituents of colon tissues, therefore, they perform the best. Similarly, EFDs features, which are based on the proposed geometric features, have also shown good classification results.

5.5. Performance analysis of hybrid feature extraction strategies

In the second set of experiments, a combinational strategy has been adopted. Manifold combinations of feature extraction

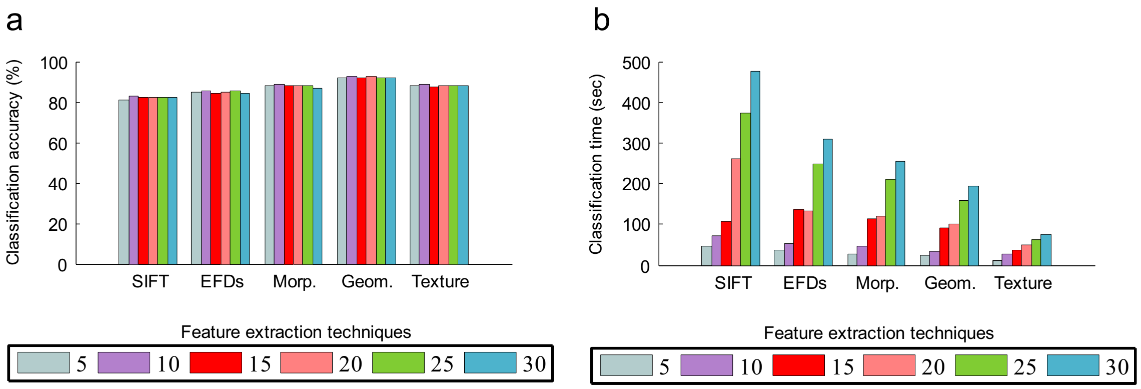


Fig. 16. Number of folds versus classification accuracy and classification time for all the feature extraction strategies.

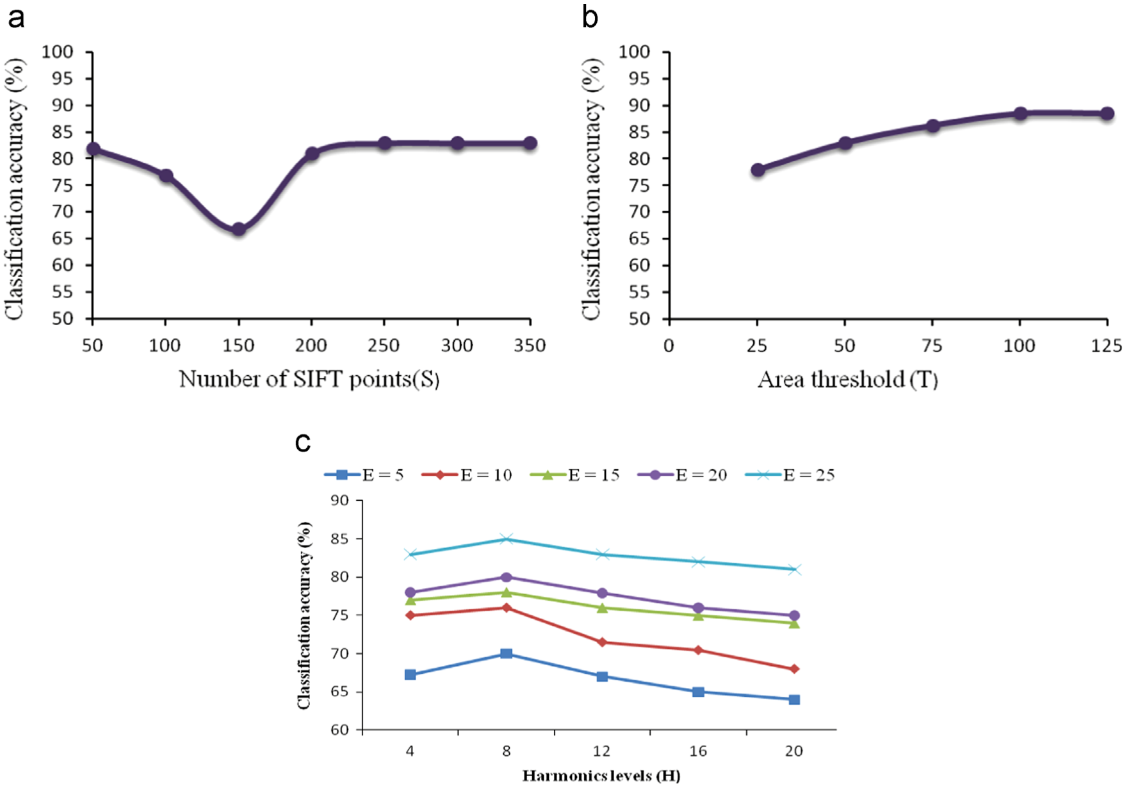


Fig. 17. Classification accuracy for different values of (a) SIFT points '*S*', (b) area threshold '*T*', (b), and (c) '*E*' and '*H*'.

Table 5
The optimal values of parameters involved in the extraction of proposed geometric features.

System parameter	Optimal values
SMJA	[9,80]
SMIA	[5,28]
RL	60
RS	40
CAT	350
OAT	1333
Membership	95

categories, each comprising two feature types, has been used for classification. Table 7 demonstrates the corresponding performance evaluation measures. Like the results of feature extraction strategies shown in Table 6, RBF kernel performs better for hybrid features as well. Therefore, Table 7 shows the performance of hybrid features achieved using RBF kernel only.

Geometric and texture features provide handsome level of accuracy no matter whether they are used separately or in combination with other feature types. Results in Table 6 and Table 7 enforce the conclusion, wherein we see that individually geometric and texture features yield maximum accuracy of 92.53% and 88.45%, respectively. However, accuracy further increases up to 94.63% when these features are combined.

Finally, we combine all the feature types to make a hybrid feature set, and achieve 96.55%, 98.28% and 98.05% classification accuracy for linear, RBF and Sigmoid kernels, respectively. The values of performance evaluation measures for hybrid feature space are given in Table 8.

Proposed technique produces quite promising results regardless of the choice of SVM kernel. This is mainly because of the diversity in the feature set encompassed by combining different feature extraction strategies. Each feature extraction strategy captures unique features from the dataset and once these features are combined; they reinforce each other, and increase accuracy upto 98.28%. An increase of 5.75% and 3.65% in accuracy is observed when compared against best performance achieved by individual feature extraction strategies and a combination of two feature extraction strategies.

The performance of the feature subsets comprising three and four feature types has also been investigated. Since, there are five individual feature extraction techniques, therefore, there are ten hybrid feature sets comprising three feature types, and five hybrid feature sets comprising four feature types. The performance of these features is superior compared to the performance of feature sets comprising two feature types, and is smaller compared to the performance of the hybrid set comprising all feature types. The best performance for the feature sets comprising three and four features types has been reported for "**Morphologic**+**Geometric**+**Texture**" and "**Morphologic**+**Geometric**+**Texture**+**EFDs**" to be 95.10% and 95.18%, respectively.

The performance of the proposed HFS-CC technique has also been analyzed in terms of the ROC curve. In this context, ROC curves of the proposed geometric features, other individual feature types, and the hybrid feature set comprising all feature types have been drawn. The ROC curves are shown in Fig. 18, which demonstrates better ROC curve of geometric feature compared to other feature extraction strategies. Secondly, the hybrid feature set comprising all feature types has the best ROC curve.

A few samples correctly classified by the proposed HFS-CC technique are shown in Fig. 19. The results show that the proposed technique has been able to correctly identify malignant images of poorly-, moderately-, and well-differentiable cancer grades.

There are a few images, which are incorrectly classified by the proposed technique. Therefore, we have strived to figure out the possible reasons of misclassification. A few normal and malignant samples misclassified by the proposed scheme are given in Fig. 20. Close observation reveals that normal images, which are misclassified, deviate from the standard geometry of normal colon tissues.

Table 6
SVM classification for individual feature extraction strategies.

Features	Accuracy	Sensitivity	Specificity	MCC	F-Score
Linear SVM					
EFDs	81.03 ± 1.32	0.80 ± 0.03	0.82 ± 0.03	0.62 ± 0.03	0.82 ± 0.01
Morphological	85.63 ± 0.36	0.91 ± 0.01	0.79 ± 0.01	0.71 ± 0.03	0.87 ± 0.02
Geometric	**91.37 ± 0.23**	**0.88 ± 0.01**	**0.95 ± 0.01**	**0.83 ± 0.02**	**0.92 ± 0.01**
Texture	86.20 ± 0.87	0.87 ± 0.01	0.85 ± 0.01	0.72 ± 0.02	0.87 ± 0.01
SIFT	78.74 ± 2.62	0.78 ± 0.04	0.79 ± 0.02	0.57 ± 0.05	0.80 ± 0.03
Sigmoid SVM					
EFDs	82.18 ± 1.54	0.80 ± 0.02	0.84 ± 0.02	0.64 ± 0.03	0.82 ± 0.02
Morphological	86.78 ± 0.78	0.92 ± 0.01	0.80 ± 0.01	0.74 ± 0.02	0.88 ± 0.01
Geometric	**91.95 ± 0.39**	**0.88 ± 0.01**	**0.96 ± 0.02**	**0.84 ± 0.02**	**0.92 ± 0.01**
Texture	87.93 ± 0.87	0.88 ± 0.01	0.88 ± 0.01	0.76 ± 0.02	0.88 ± 0.01
SIFT	81.03 ± 0.77	0.82 ± 0.02	0.80 ± 0.02	0.62 ± 0.03	0.82 ± 0.03
RBF SVM					
EFDs	85.06 ± 1.09	0.85 ± 0.03	0.85 ± 0.02	0.70 ± 0.02	0.86 ± 0.01
Morphological	88.50 ± 0.78	0.92 ± 0.02	0.84 ± 0.02	0.77 ± 0.01	0.89 ± 0.01
Geometric	**92.53 ± 0.15**	**0.87 ± 0.01**	**0.99 ± 0.01**	**0.86 ± 0.02**	**0.92 ± 0.01**
Texture	88.45 ± 2.19	0.84 ± 0.02	0.92 ± 0.02	0.77 ± 0.02	0.87 ± 0.01
SIFT	82.76 ± 2.19	0.86 ± 0.01	0.79 ± 0.01	0.65 ± 0.01	0.84 ± 0.03

Table 7
SVM classification for combination of feature extraction strategies using RBF kernel of SVM.

Features	Accuracy	Sensitivity	Specificity	MCC	F-Score
EFDs-Morphological	88.58 ± 0.98	0.88 ± 0.01	0.90 ± 0.01	0.80 ± 0.02	0.89 ± 0.01
EFDs-Geometric	92.88 ± 1.11	0.89 ± 0.02	0.95 ± 0.01	0.85 ± 0.03	0.92 ± 0.01
EFDs-Texture	89.82 ± 0.67	0.89 ± 0.01	0.91 ± 0.01	0.80 ± 0.01	0.89 ± 0.01
EFDs-SIFT	83.79 ± 1.40	0.89 ± 0.02	0.79 ± 0.02	0.68 ± 0.03	0.84 ± 0.01
Morphological-Geometric	94.57 ± 0.63	0.92 ± 0.01	0.90 ± 0.01	0.90 ± 0.01	0.93 ± 0.01
Morphological-Texture	92.36 ± 0.88	0.91 ± 0.01	0.94 ± 0.01	0.85 ± 0.01	0.93 ± 0.01
Morphological-SIFT	91.23 ± 0.33	0.90 ± 0.01	0.92 ± 0.01	0.82 ± 0.01	0.91 ± 0.01
Geometric-Texture	**94.63 ± 0.82**	**0.92 ± 0.01**	**0.95 ± 0.01**	**0.90 ± 0.02**	**0.92 ± 0.01**
Geometric-SIFT	93.39 ± 0.68	0.93 ± 0.01	0.94 ± 0.01	0.87 ± 0.01	0.93 ± 0.01
Texture-SIFT	90.80 ± 0.77	0.85 ± 0.02	0.96 ± 0.01	0.82 ± 0.01	0.90 ± 0.01

Table 8
SVM classification for hybrid feature space.

SVM kernels	Accuracy	Sensitivity	Specificity	MCC	F-Score
Linear	96.55 ± 0.25	0.97 ± 0.01	0.96 ± 0.01	0.93 ± 0.01	0.97 ± 0.01
Sigmoid	98.05 ± 0.57	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.01	0.98 ± 0.01
RBF	**98.28 ± 0.36**	**0.99 ± 0.01**	**0.98 ± 0.02**	**0.97 ± 0.01**	**0.98 ± 0.01**

These images actually depict pre-cancerous stage in which deformation process is about to begin. Similarly, a malignant image which is misclassified by the proposed HFS-CC technique is shown in Fig 20(a3). This image lies at the boundary between normal and malignant tissues. Therefore, system finds difficulty in accurately identifying this image as malignant one.

5.6. *Computational time requirements of the proposed HFS-CC technique*

It is always desirable to measure the effectiveness of a proposed technique in terms of its CPU time requirement. In this context, we have calculated the CPU time taken by different feature extraction strategies, and also the time taken by different classifiers for classification of individual and hybrid features. The sizes of the individual and hybrid feature sets, and the feature extraction time for each feature type are given in Fig. 21(a) and (b), respectively. The feature extraction time of hybrid feature set is mere addition of the extraction time of individual feature sets.

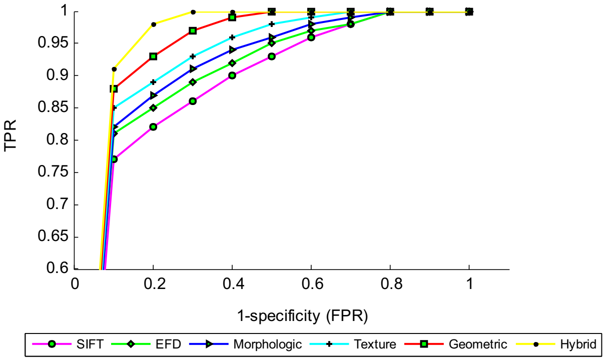
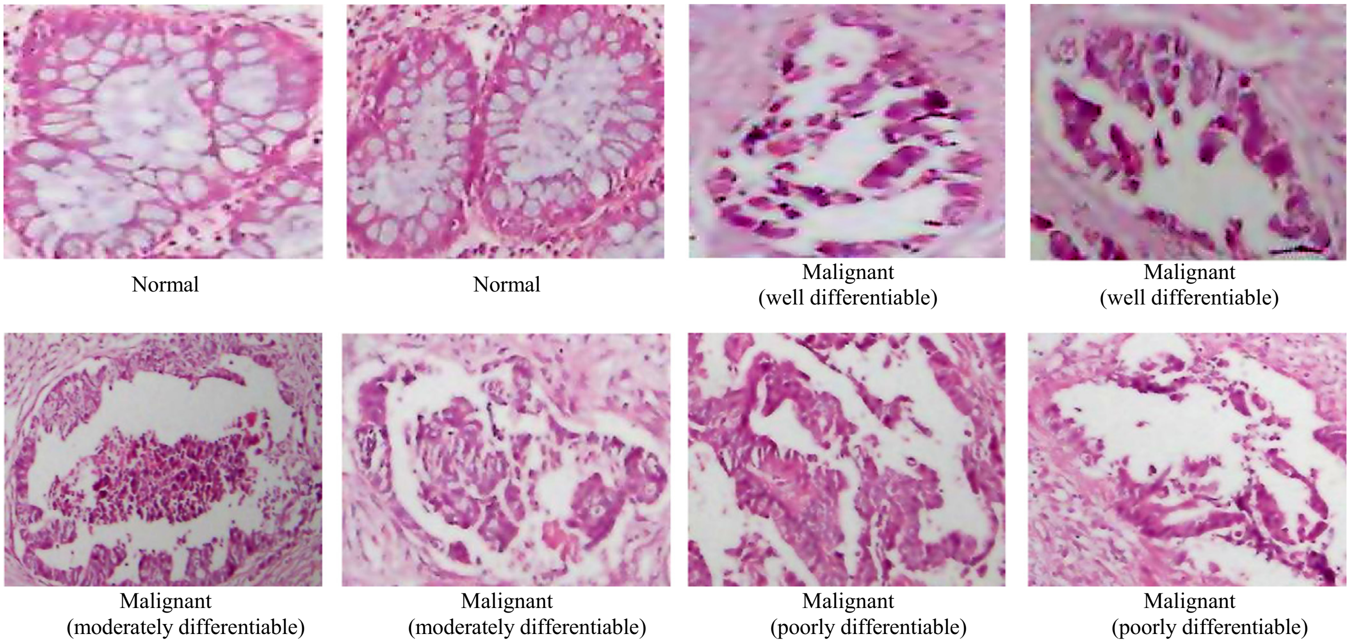


Fig. 18. ROC curves for individual and hybrid feature sets.

Fig. 21(b) demonstrates that the proposed HFS-CC technique is computationally tractable, because the extraction of various feature types from an image takes quite small time. Even the extraction of hybrid feature set, which is the sum of individual feature extraction times, is only 12.08 s. Further, the extraction of geometric features is time-consuming compared to other feature types. Reason behind the fact is that geometric features are based on objects. These objects are located by searching multiple simulated ellipses in the whole image, thereby increasing the feature extraction time. But the increase in computational time may be justified by the better performance of geometric features compared to other feature types.

Please cite this article as: S. Rathore, et al., Automated colon cancer detection using hybrid of novel geometric features and some traditional features, Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.03.004


Normal

Normal

Malignant (well differentiable)

Malignant (well differentiable)

Malignant (moderately differentiable)

Malignant (moderately differentiable)

Malignant (poorly differentiable)

Malignant (poorly differentiable)

Fig. 19. Examples of the normal, poorly-, moderately-, and well differentiable malignant colon tissues, which are correctly classified by the proposed HFS-CC system.

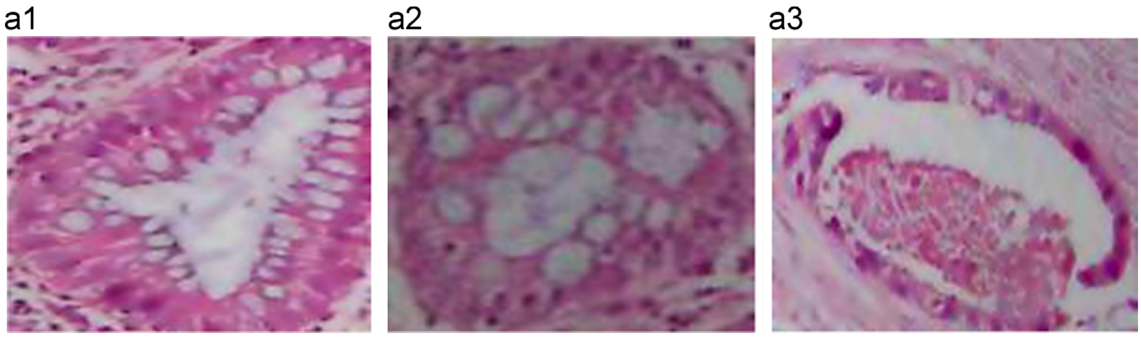


Fig. 20. Misclassified ($a_1 - a_2$) normal and (a_3) well-differentiated malignant image by the proposed HFS-CC using RBF SVM.

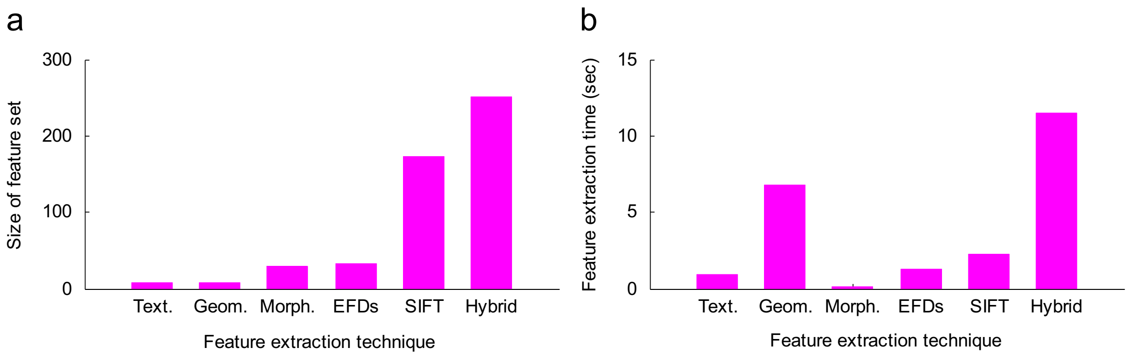


Fig. 21. (a) Size, and (b) feature extraction time of various feature sets.

Fig. 22 demonstrates the classification time of different feature sets taken by various SVM classifiers. We observe from the results that classification time depends upon the size of the feature vector. Hence, the classification of hybrid feature set takes considerably large time compared to individual feature sets.

5.7. Performance of the proposed HFS-CC technique on separate training and test data

This section separately investigates the training and testing performance of the proposed system. In this context, overall dataset has been randomly divided into training and test sets. The training set consists of 70% of the overall dataset, which is 122 images (56 normal, and 66 malignant). The test set comprises 30% of the dataset, which is 56 images (26 normal, and 26 malignant). The images in the test data are not used in parameter estimation at all, and are totally unseen for the model trained on the training data.

In the training phase, the optimal values of system's parameters are determined through 10-fold cross-validation, and a model is trained on these parameters. The proposed/used feature selection methodologies have some parameters. Additionally, there are some parameters of linear, RBF and sigmoid kernels of SVM. In the training phase, we consider all possible combinations of these parameters as candidate parameter sets. Using 10-fold cross validation on the training set, we select the optimal values of these parameters independently for different feature selection methodologies.

In the testing phase, the test data is applied on the model obtained in the training phase, and classes of the samples are determined. The test data is totally unseen for the model trained during training. The training and test accuracy for different feature selection methodologies is given in Table 9.

The results in Table 9 are almost similar to those obtained in Section 5.4. The results show that all the feature types yield good classification results, but the proposed geometric features outclass other feature types, and yield better classification.

The training and testing performance for hybrid feature set comprising all feature types is given in Table 10. The results show that the proposed hybrid feature set has promising classification results. The classification accuracies of 99.18% and 98.07% in the training and test phases show that the proposed hybrid feature set discriminates normal and malignant colon tissues quite effectively.

5.8. Performance comparison of HFS-CC with existing schemes

The performance of the proposed HFS-CC system has been compared with previously proposed approaches of colon biopsy image classification. In this context, six techniques [11–16] have been selected from the contemporary literature for comparison. We have implemented these techniques in Matlab, and evaluated classification performance measures on the dataset described in Section 5.1. In order to obtain a fair comparison with HFS-CC, we have used optimal values of the parameters used in these techniques.

Table 11 reveals better performance of HFS-CC over others in terms of most of the performance evaluation measures. This is because of the fact that some of the previous techniques [11–14] take general image features for classification such as entropy, correlation, and inverse difference moment. These features represent general image texture but do not capture any problem related information such as information about tissue organization. On the other hand, proposed scheme uses rich hybrid feature vector wherein each individual feature category captures problem specific knowledge. When individual feature vectors are combined, they all reinforce each other and produce quite superior classification results. This work also has some advantages over the authors' previously proposed CBIC technique [15]. In CBIC, a majority voting based ensemble of three individual SVM classifiers has been developed for classification. The parameter optimization and the classification time of ensemble are equal to the sum of the time consumed by individual classifiers. Furthermore, an overhead of majority voting also exists in CBIC. On the other hand, only one classifier (RBF that gives better results) has been used in this work that not only produces better results on training data and comparable results on test data, but also consumes one third of the time compared to CBIC in parameter optimization and classification. Furthermore, variants of statistical moments and Haralick

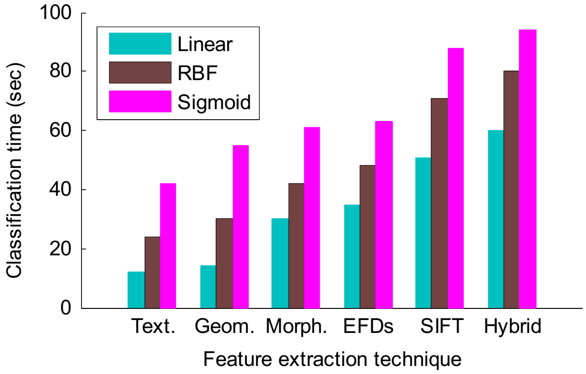


Fig. 22. Classification time of different feature sets by various SVM classifiers.

Table 10
Training and testing performance of linear, RBF and sigmoid SVM kernels for hybrid feature extraction strategy.

SVM Kernels	Training performance	Test performance				
	Accuracy	Accuracy	Sensitivity	Specificity	MCC	F-Score
Linear	96.72	94.23	0.96	0.92	0.88	0.94
Sigmoid	97.54	96.15	0.96	0.96	0.92	0.96
RBF	**99.18**	98.07	1.00	0.96	0.96	0.98

Table 9
Training and testing performance of linear, RBF and sigmoid SVM kernels for individual feature extraction strategies.

Features	Training performance	Testing performance				
	Accuracy	Accuracy	Sensitivity	Specificity	MCC	F-score
Linear SVM						
EFDs	82.79	82.69	0.85	0.81	0.65	0.83
Morphological	85.24	84.61	0.88	0.81	0.69	0.85
Geometric	**90.98**	**90.38**	**0.88**	**0.92**	**0.81**	**0.90**
Texture	86.07	84.61	0.77	0.92	0.70	0.83
SIFT	77.87	76.92	0.96	0.58	0.58	0.81
Sigmoid SVM						
EFDs	83.61	82.69	0.92	0.73	0.67	0.84
Morphological	84.43	84.61	0.85	0.85	0.69	0.85
Geometric	**91.80**	**92.30**	**0.92**	**0.92**	**0.85**	**0.92**
Texture	86.89	86.53	0.92	0.81	0.74	0.87
SIFT	83.61	82.69	0.81	0.85	0.65	0.82
RBF SVM						
EFDs	86.07	86.54	0.81	0.92	0.74	0.86
Morphological	90.16	88.46	0.85	0.92	0.77	0.88
Geometric	**92.62**	**94.23**	**0.88**	**1.00**	**0.89**	**0.94**
Texture	87.70	88.46	0.96	0.81	0.78	0.89
SIFT	83.61	84.62	0.73	0.96	0.71	0.83

Table 11
Comparison of HFS-CC with some existing schemes.

Technique	Accuracy	Sensitivity	Specificity	MCC	F-score
Esgiar et al. [11]	86.00 ± 1.32	0.89 ± 0.09	0.95 ± 0.04	0.54 ± 0.07	0.84 ± 0.12
Esgiar et al. [12]	73.18 ± 1.78	0.82 ± 0.02	0.67 ± 0.02	0.48 ± 0.03	0.71 ± 0.02
Masood et al. [13]	87.93 ± 1.52	0.91 ± 0.02	0.83 ± 0.01	0.78 ± 0.02	0.88 ± 0.02
Masood et al. [14]	89.88 ± 1.32	0.92 ± 0.03	0.88 ± 0.02	0.79 ± 0.03	0.91 ± 0.03
Altunbay et al. [16]	91.95 ± 0.95	0.91 ± 0.02	0.93 ± 0.02	0.84 ± 0.03	0.92 ± 0.01
Rathore et al. [15]	98.85 ± 0.51	0.98 ± 0.01	1.00 ± 0.00	0.98 ± 0.01	0.99 ± 0.00
HFS-CC (test)	98.07 ± 0.23	1.00 ± 0.00	0.96 ± 0.01	0.96 ± 0.01	0.98 ± 0.01
HFS-CC (training)	99.18 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.00

features have been used in CBIC. On the other hand, novel geometric features have been proposed for classification of colon biopsy images in this research work. These features have good classification results, and may further be improved to produce more accurate classification results. These features may not only be helpful for the researchers working in the field of colon cancer, but also for those working with other cancer types.

6. Conclusion

In this research study, a novel classification technique HFS-CC is proposed for predicting cancer in colon tissues. In the proposed scheme, several features such as morphological, texture, EFDs and SIFT are extracted from colon biopsy images. Further, a novel feature type is proposed that quantifies the knowledge about the size and the spatial distribution of various cytological constituents of colon tissues for developing a feature vector to be used in the classification. Further, traditional features are combined with the proposed geometric features to form a hybrid feature vector, which is then used in different kernels based SVM classification. Working with colon biopsy images, 99.18% training and 98.07% test classification precision has been observed. Proposed technique has also been compared with various existing colon cancer detection techniques, and a significant increase in classification accuracy has been observed. Results demonstrate that hybrid and rich feature space certainly improves the classification performance compared to the performance achieved by using individual features. This research study can be extended into multiple directions. First possibility is to use supervise/unsupervised techniques for assigning quantitative cancer grades to already classified malignant samples. Second, the proposed scheme can be tested on Immuno Histochemically stained biopsies to measure its effectiveness. Third, some valuable color based features can also be extracted from colon biopsy images.

Acknowledgement

This work is supported by the Higher Education Commission of Pakistan (HEC) under indigenous PhD scholarship program as per award no. 117-7931-Eg7-037. We are thankful to the histopathologists, who actively remain involved in data and ground truth preparation.

References

[1] A.B. Tosun, M. Kandemir, C. Sokmensuer, C.G. Demir, Object-oriented texture analysis for the unsupervised segmentation of biopsy images, J. Pattern Recognit. 42 (2009) 1104–1112.
[2] G.D. Thomas, M.F. Dixon, N.C. Smeeton, N.S. Williams, Observer variation in the histological grading of rectal carcinoma, J. Clin. Pathol. 36 (1983) 385–391.
[3] A. Andrion, C. Magnani, P.G. Betta, A. Donna, F. Mollo, M. Scelsi, P. Bernardi, M Botta, B. Terracini, Malignant mesothelioma of the pleura: inter observer variability, J. Clin. Pathol. 48 (1995) 856–860.
[4] S. Rathore, M. Hussain, A. Ali, A. Khan, A recent survey on colon cancer detection techniques, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2013) 545–563.
[5] K. Rajpoot, N. Rajpoot, SVM optimization for hyperspectral colon tissue cell classification, in: Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI), Lecture Notes in Computer Science 3217 Saint-Malo, France, 2004, pp. 829–837.
[6] R.J. Cassidy, J. Berger, K. Lee, M. Maggioni, R.R. Coifman, Analysis of hyperspectral colon tissue images using vocal synthesis models, in: Proceedings of 38th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, California, 2004, pp. 1611–1615.
[7] E.T. Venkatesh, P. Thangaraj, S. Chitra, An improved neural approach for malignant and normal colon tissue classification from Oligonucleotide arrays, Eur. J.Sci. Res. 54 (2011) 159–164.
[8] M. Tong, K.H. Liu, C. Xu, W. Ju, An ensemble of SVM classifiers based on gene pairs, Comput. Biol. Med. 43 (2013) 729–737.
[9] X. Li, X. Li, M. Lei, D. Wang, J. Lin, Detection of colon cancer by laser induced fluorescence and raman spectroscopy, Proceedings of 27th International Conference of Engineering in Medicine and Biology Society 2005. 6961–6964.
[10] Z. Huang, Laser-induced auto fluorescence microscopy of normal and tumor human colonic tissue, Int. J. Oncol. 24 (2004) 59–63.
[11] A.N. Esgiar, R.N.G. Naguib, B.S. Sharif, M.K. Bennett, A. Murray, Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa, IEEE Trans. Inform. Technol. Biomed. 2 (1998) 197–203.
[12] A.N. Esgiar, R.N.G. Naguib, B.S. Sahrif, M.K. Bennett, Fractal analysis in the detection of colonic cancer images, IEEE Trans. Inform. Technol. Biomed. 6 (2002) 54–58.
[13] K. Masood, N. Rajpoot, H. Qureshi, K. Rajpoot, Co-occurrence and morphological analysis for colon tissue biopsy classification, Proceedings of 4th International Workshop on Frontiers of Information Technology 2006. 211–216.
[14] K. Masood, N. Rajpoot, Texture based classification of hyperspectral colon biopsy samples using CLBP, Proceedings of International Symposium on Biomedical Imaging: From Nano to MacroBoston 2009. 1011–1014.
[15] S. Rathore, M. Hussain, M.A. Iftikhar, A. Jalil, Ensemble classification of colon biopsy images based on information rich hybrid features, Comput. Biol. Med. 47 (2013) 76–92.
[16] D. Altunbay, C.D. Altunbay, C. Cigir, C. Sokmensuer, C.G. Demir, Color graphs for automated cancer diagnosis and grading, IEEE Trans. Biomed. Eng. 57 (2010) 665–674.
[17] E. Ozdemir, C.G. Demir, A hybrid classification model for digital pathology using structural and statistical pattern recognition, IEEE Trans. Med. Imag. 32 (2013).
[18] A.B. Tosun, C. Gunduz-Demir, Graph run-length matrices for histopathological image segmentation, IEEE Trans. Med. Imag. 30 (2011) 721–732.
[19] K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Trans. Inform. Theory 21 (1975) 32–40.
[20] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 790–799.
[21] D. Comaniciu, P. Meer, Robust analysis of feature spaces: color image segmentation, Proceedings of the International Conference on Computer Vision and Pattern Recognition 1997750–755.
[22] M. Masseroli, A. Bollea, G. Forloni, Quantitative morphology and shape classification of neurons by computerized image analysis, Comput. Methods Prog. Biomed. 41 (1993) 89–99.
[23] D. Welfer, J. Scharcanski, D.R. Marinho, Fovea center detection based on the retina anatomy and mathematical morphology, Comput. Methods Prog. Biomed. 104 (2011) 397–409.
[24] V. Naranjo, R. Llorens, M. Alcaniz, F. Lopez-Mir, Metal artifact reduction in dental CT images using polar mathematical morphology, Comput. Methods Prog. Biomed. 102 (2011) 64–74.
[25] Y.M. Li, X.P. Zeng, A new strategy for urinary sediment segmentation based on wavelet, morphology and combination method, Comput. Methods Prog. Biomed. 84 (2006) 162–173.
[26] R.C. Gonzalez, R.E. Woods, Digital Image Processing, Prentice Hall, 2002.
[27] D.S. Guru, Y.H. Sharath, S. Manjunath, Texture features and KNN in classification of flower images, Proceedings of IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition (2010) 21–29.

[28] S.G. Mougiakakou, I. Valavanis, K.S. Nikita, A. Nikita, D. Kelekis, Characterization of CT liver images based on texture features and a multiple neural network classification scheme, Proceedings of 25th Annual International Conference on Engineering in Medicine and Biology Society 2003. 1287–1290.

[29] M.E. Mavroforakis, H.V. Georgiou, D. Cavouras, N. Dimitropoulos, S. Theodoridis, Mammographic mass classification using textural features and descriptive diagnostic data, Proceedings of 14th International Conference on Digital Signal Processing 2002. 461–464.

[30] D.G. Lowe, Dsitinctive image features from scale invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[31] M. Brown, D.G. Lowe, Recognizing panoramas, Proceedings of 9th International Conference on Computer Vision 2003. 1218–1225.

[32] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kewade, B.L. Lu, Person specific SIFT features for face recognition, Proceedings of International Conference on Acuostics, Speech and Signal Processing 2007. 593–596.

[33] M. Bicego, A. Lagorio, E. Grosso, M. Tistarelli, On the use of SIFT features for face authentication, Proceedings of International Conference on Computer Vision and Pattern Recognition Workshop 2006. 35.

[34] D.R. Kisku, A. Rattani, E. Grosso, M. Tistarelli, Face identification by SIFT-based complete graph topology, Proceedings of Automatic Identification Advanced Technologies 2007. 63–68.

[35] S. Fazli, H.M. Pour, H. Bouzari, Particle filter based object tracking with SIFT and color feature, Proceedings of 2nd International Conference on Machine Vision 2009. 89–93.

[36] Fast SIFT Image Features Library, 2012.

[37] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and minredundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.

[38] F.P. Kuhl, C.R. Giardina, Elliptic fourier features of a closed contour, Comput. Graph. Image Process. 18 (1982) 236–258.

[39] L.P. Nicoli, G.C. Anagnostopoulos, Shape-based recognition of targets in synthetic aperture radar images using elliptical Fourier descriptors, Proceedings of SPIE (2008).

[40] T. Taxt, U. Bergen, K.W. Bjerde, Classification of hand written vector symbols using elliptic fourier descriptos, Proceedings of the 12th International Conference on Computer Vision and Image Processing 1994. 123–128.

[41] A. Subasi, Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders, Comput. Biol. Med. 43 (2013) 576–586.

[42] P.D. Andrzej, M. Wierzbowski, K. Tomczykiewicz, Multiresolution MUAPs decomposition and SVM-based analysis in the classification of neuromuscular disorders, Comput. Biol. Med. 107 (2012) 393–403.

[43] W.B. Sampaioa, E.M. Diniza, A.C. Silvaa, A.C. Paivaa, M. Gattassb, Detection of masses in mammogram images using CNN, geostatistic functions and SVM, Comput. Biol. Med. 41 (2011) 653–664.

[44] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, London, 2005.

[45] C.W. Hsu, C.C. Chang, C.J. Lin, A practical guide to support vector machines, Department of Computer Science & Information Engineering, National Taiwan University, 2003.

[46] F.H.F. Leung, H.K. Lam, S.H. Ling, P.K.S. Tam, Tuning of the structure and parameters of a neural network using an improved genetic algorithm, IEEE Trans. Neural Netw. 14 (2003) 79–88.

[47] D.R. ElShafie, N. Kharma, R. Ward, Parameter optimization of an embedded watermark using a genetic algorithm, Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing, ISCCSP 2008. 1263–1267.

Please cite this article as: S. Rathore, et al., Automated colon cancer detection using hybrid of novel geometric features and some traditional features, Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.03.004