COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 121 (2015) 92–108

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Novel structural descriptors for automated colon cancer detection and grading

Saima Rathore[a,b,*], *Mutawarra Hussain*[a], *Muhammad Aksam Iftikhar*[a,c], *Abdul Jalil*[a]

[a] *DCIS, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan*
[b] *DCS&IT, University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir, Pakistan*
[c] *Comsats Institute of Information Technology, Lahore, Pakistan*

ARTICLE INFO

Article history:
Received 26 December 2014
Received in revised form
25 May 2015
Accepted 27 May 2015

Keywords:
Colon cancer
Colon cancer detection
Colon cancer grading
Colon classification

ABSTRACT

The histopathological examination of tissue specimens is necessary for the diagnosis and grading of colon cancer. However, the process is subjective and leads to significant inter/intra observer variation in diagnosis as it mainly relies on the visual assessment of histopathologists. Therefore, a reliable computer-aided technique, which can automatically classify normal and malignant colon samples, and determine grades of malignant samples, is required. In this paper, we propose a novel colon cancer diagnostic (CCD) system, which initially classifies colon biopsy images into normal and malignant classes, and then automatically determines the grades of colon cancer for malignant images. To this end, various novel structural descriptors, which mathematically model and quantify the variation among the structure of normal colon tissues and malignant tissues of various cancer grades, have been employed. Radial basis function (RBF) kernel of support vector machines (SVM) has been employed as classifier in order to classify/grade colon samples based on these descriptors. The proposed system has been tested on 92 malignant and 82 normal colon biopsy images. The classification performance has been measured in terms of various performance measures, and quite promising performance has been observed. Compared with previous techniques, the proposed system has demonstrated better cancer detection (classification accuracy = 95.40%) and grading (classification accuracy = 93.47%) capability. Therefore, the proposed CCD system can provide a reliable second opinion to the histopathologists.

1. Introduction

Colon cancer is one of the leading causes of cancer related deaths in modern and industrialized world. About half a million people die every year worldwide due to colon cancer [1]. Primary reason of colon cancer is chain smoking, but there are some other reasons of colon cancer such as family history of colon cancer, increasing age, and unbalanced diet such as diets with low consumption of fruits/vegetables and heavy consumption of meat [2].

The conventional method of colon cancer diagnosis is microscopic analysis of colon biopsy samples. In such an examination, histopathologists analyze the biopsy samples under microscope, and diagnose the tissue as normal/malignant based on the morphology of tissues. Further,

* *Corresponding author at*: DCIS, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan. Tel.: +92 51 2207381x3102.
E-mail address: saimarathore_2k6@yahoo.com (S. Rathore).

http://dx.doi.org/10.1016/j.cmpb.2015.05.008
0169-2607/

histopathologists assign quantitative cancer grades depending upon the morphology of malignant tissues they observe under microscope. However, the manual process of colon cancer diagnosis has a few limitations. First, it consumes valuable time of the histopathologists as they have to examine many images per day. Second, the process is subjective, and may lead to inter- and intra-observer variability in diagnosis due to workload and experience of histopathologists [3,4]. Therefore, a computer-aided diagnostic system, which could accurately identify normal and malignant samples, and quantify the cancer grade as well, is required.

In the past decade, a few computer-aided diagnostic systems have been proposed for automatic detection of normal and malignant colon biopsy images. These techniques exploit the textural changes in normal and malignant colon biopsy images, and have been summarized in a recent survey [5]. The texture analysis of colon biopsy images is characterized by the extraction of discerning features from the images. The extracted features are then used as input to different classifiers for identifying normal and malignant images. For instance, Esgiar et al. calculated texture features of contrast, entropy, angular second moment, dissimilarity, inverse difference moment and correlation from gray-level co-occurrence matrix (GLCM) of the input colon biopsy images [6]. They achieved 90.20% classification accuracy by employing linear discriminate analysis (LDA) and K-nearest neighbor (KNN) classifiers. In another work, Esgiar et al. combined features of entropy and correlation with image fractal dimensions [7], and obtained 94.10% classification accuracy with the same set of classifiers i.e. KNN and LDA. Later, Masood et al. employed GLCM based texture features of energy, inertia and local homogeneity, and morphological features of shape, size and orientation in order to classify normal and malignant colon biopsy images. They employed third degree polynomial kernel of support vector machines (SVM) for classification, and achieved an accuracy of 84% and 90%, respectively, by using morphological and texture features [8]. Masood et al. further extended their work, and employed circular local binary patterns in order to classify colon biopsy images [9]. They employed Gaussian kernel of SVM for classification, and obtained classification success rate of 90%.

Similarly, a few cancer grading techniques have also been proposed in the past. For example, Altunbay et al. proposed a textural features based technique for classifying colon samples into normal and malignant (low grade and high grade) categories [10]. They constructed a graph on different objects, obtained by using circle fit algorithm [11] on the white, pink and purple clusters of the image. A few structural features such as degree, average clustering coefficient, and diameter are computed from the graphs, and are used to classify given images by using linear SVM.

Ozdemir and Demir also presented an automatic colon cancer detection and grading technique [12]. In this work, a few normal colon tissues are manually selected among large sized colon tissue images, and query graph is generated separately on each selected tissue. Later, reference graph is generated on each test colon biopsy image. The query and the reference graphs are generated using the same way as done in [11]. Further, multiple key regions are located in the reference image that are most similar to a normal biological structure by searching each query graph over the entire reference graph of test image. The similarity is measured in terms of graph edit distance. The graph edit distance actually quantifies the dissimilarity in source and the destination graph by calculating the minimum cost of edit operations that should be applied on source to transform it into destination graph. The basic idea behind this approach is that query graphs of normal glands will be similar to the graphs of selected key regions of normal colon biopsy images compared to those of malignant colon biopsy images. Later, graph edit distance, and some other statistical features computed from these key-regions are given as input to SVM classifier for classification of test images into normal, and malignant tissues of low and high grade cancer. They achieved 92.21% classification accuracy on a colon biopsy image based dataset. But, this technique is computationally expensive due to heavy processing involved in matching query graphs and key regions of reference graphs.

The techniques mentioned in the previous paragraphs have a few limitations. First, the graph based techniques [10,12] are computationally expensive. Generating graphs for extracting these features is not computationally expensive. Rather, these techniques consume large time in fitting circles in the three clusters of colon biopsy images for determining graph nodes, which is a pre-requisite step of graph generation. Further, feature extraction from graphs is also expensive in some of these techniques such as Ozdemir and Demir [12], wherein comparison of query graphs and key regions of reference graphs is expensive. Second major set of techniques for feature extraction [6–9] are those that exploit the texture variation in normal and malignant images. These texture features are general in nature and do not consider the specific pathological variation between normal and malignant colon tissues. For example, colon cancer causes the lumen tissue in a malignant image to turn from near-elliptic to irregular shape. Such pathological variation in structures of tissues is not encoded in the features proposed in literature for colon cancer classification/grading. Therefore, a computationally tractable computer-aided diagnostic technique is required, which could exploit the background knowledge specifically about pathological structure of normal and malignant colon tissues into the classification process.

In this paper, we propose a novel colon cancer diagnosis (CCD) system, which is capable of automatic colon cancer detection, and its classification into various cancer grades. Unlike previously proposed methods, which capture information about the general texture present in colon biopsy images, the proposed method incorporates the background pathological information about the morphology of normal and malignant tissues into the classification and grading process. To this end, some novel structural features, which exploit the shape, convexity, concavity, circularity and area based characteristics of lumen for detection and grading of colon cancer, have been proposed. A data set comprising 192 images has been used for validating the proposed CCD system. Training and testing data has been formulated using Jack-knife 10-fold cross-validation, and radial basis function (RBF) kernel of SVM classifier has been employed for both cancer detection and cancer grading. The proposed features have been proved to yield better results compared to general texture based features in the experimental section.

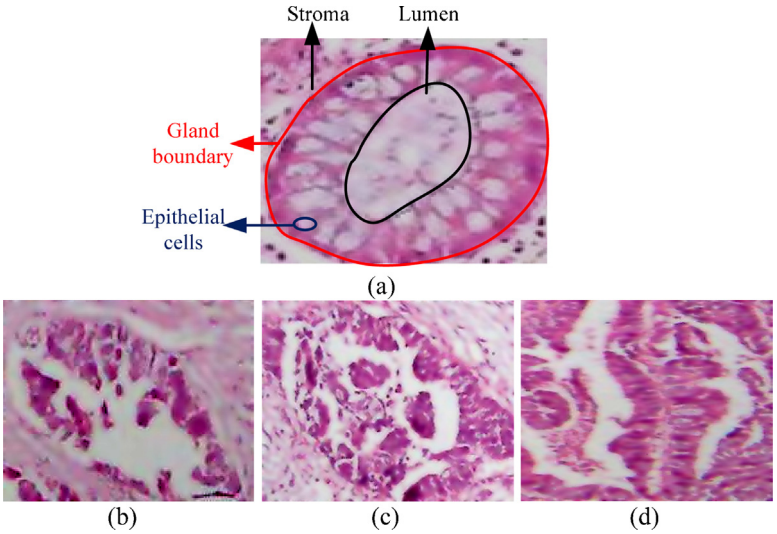

Fig. 1 – Organizational structure of (a) normal colon tissue, and malignant colon tissues of (b) well-, (c) moderate-, and (d) poor-differentiable colon cancer grades as observed under microscope.

The rest of this article is structured as follows. Section 2 describes the morphology of normal and malignant colon tissues. Section 3 illustrates the proposed technique, and the novel structural descriptors in detail. Section 4 presents performance measures, which have been used for evaluation purposes. Section 5 provides experimental setup, results and the associated discussion. Section 6 concludes this research work.

2. Morphology of normal and malignant colon tissues

A normal colon tissue has three main constituents, namely epithelial cells, non-epithelial cells, and lumen. The detailed organizational structure of a normal colon tissue is shown in Fig. 1(a), wherein we see that the said constituents possess a regular organizational structure, and are very much aligned. Epithelial cells surround lumen and form glandular structure, whereas non-epithelial cells (stroma) lie outside these structures.

Normal colon tissues have well-defined structures such as elliptic shaped epithelial cells and lumen. However, the standard geometry of normal colon tissues is disturbed considerably if the patient is suffering from cancer. In malignant colon tissues, the constituents of tissues mix together, thereby diminishing the boundaries and introducing deformation. Further the deformation introduced by cancer increases as the cancer grade progresses. The grade of colon cancer actually quantifies the differentiability level of malignant cells. There are three grades of colon cancer, namely well-, moderate-, and poor-differentiable. The 'well differentiable' is the slowly progressing grade of colon cancer in which malignant cells are nearly similar to the normal ones. In 'moderately differentiable' grade of colon cancer, malignant cells are differentiable from normal cells, and cancer progresses at moderate speed. The third grade of colon cancer is 'poor differentiable'. In this particular grade, malignant cells are totally different from the normal ones, and spread at very high rate. Fig. 1(b)–(d) presents microscopic images of well-, moderate-, and poor-differentiable colon samples, respectively.

3. Proposed system

The proposed system comprises three main phases, namely (1) pre-processing, (2) cancer detection, and (3) cancer grading. Fig. 2 portrays the three phases of the proposed system. In the pre-processing phase, a few pre-processing steps are applied on image database in order to make the images suitable for subsequent processing. In the cancer detection phase, a few novel descriptors, which exploit the structural differences between morphology of normal and malignant colon tissues, are extracted from the pre-processed colon biopsy images. These descriptors are then employed for classification of images into normal and malignant categories. The next phase deals with the grading of malignant images. In this phase, the structural changes happening in colon biopsy images due to progressing cancer grade are exploited, and a few descriptors are extracted exploiting these structural changes. These descriptors are then used for the identification of various cancer grades. Thus, the proposed CCD system may be used by histopathologists as a comprehensive tool for detection and grading of colon cancer. Different phases of the proposed system are discussed in detail in the following text.

3.1. *Pre-processing phase*

The pre-processing step is generally performed in order to make the images suitable for next phases especially for the feature extraction phase. In this research work, *K*-Means algorithm is applied as a pre-processing strategy in order to

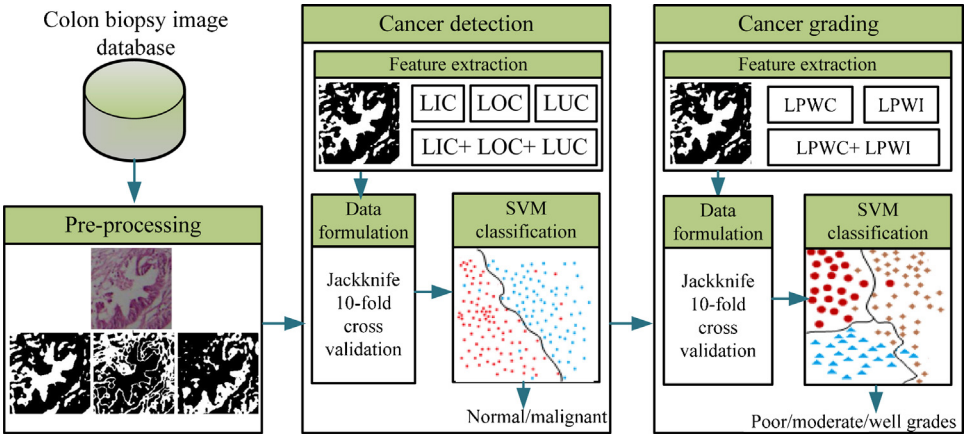

Fig. 2 – Schematic overview of the proposed CCD system comprising preprocessing, cancer detection, and cancer grading phases.

divide colon biopsy images into constituent clusters. Colon biopsy images are usually characterized by pink-colored connecting tissues, purple-colored nucli, and white-colored epithelial cells and lumen. Therefore, *K*-Means algorithm with $K=3$ is applied on color intensities of pixels in order to divide image pixels into three clusters. *K*-Means is a non-parametric statistically iterative method, originally developed by Fukunaga et al. for estimation of gradients of a density function [13], and has extensive use in computer vision for image segmentation [14,15] and visual object tracking [16]. *K*-Means is based on a simple and straightforward concept. It randomly picks large number of image pixels as representatives of cluster centers. A hypothesized multidimensional ellipsoid is centered on cluster center, and the cluster center is moved to the mean of the data lying inside ellipsoid. Mean is iteratively calculated, and cluster centers are moved accordingly until there is no significant change in mean value. Adjacent and similar regions are merged during iterations, and number of final clusters may be much smaller than the number of clusters in the start. The output of applying *K*-Means on a colon biopsy image is shown in Fig. 3.

3.2. *Cancer detection phase*

The purpose of cancer detection phase is to identify normal and malignant colon biopsy images. In this phase, some novel structural descriptors are extracted from the white cluster of colon biopsy images. The white cluster corresponds to epithelial cells and lumen, which undergo maximal variation in case of malignancy. Thus, extracting structural features from the white cluster captures the underlying difference between the structures of normal and malignant images. The descriptor values are then formulated, and images are classified into normal and malignant categories based on the descriptors. Three novel structural descriptors, namely, lumen inner concavity (LIC), lumen outer convexity (LOC), and lumen circularity (LUC) have been proposed for cancer detection. LIC and LOC have been explained in Section 3.2.1, whereas LUC has been explained in Section 3.2.2.

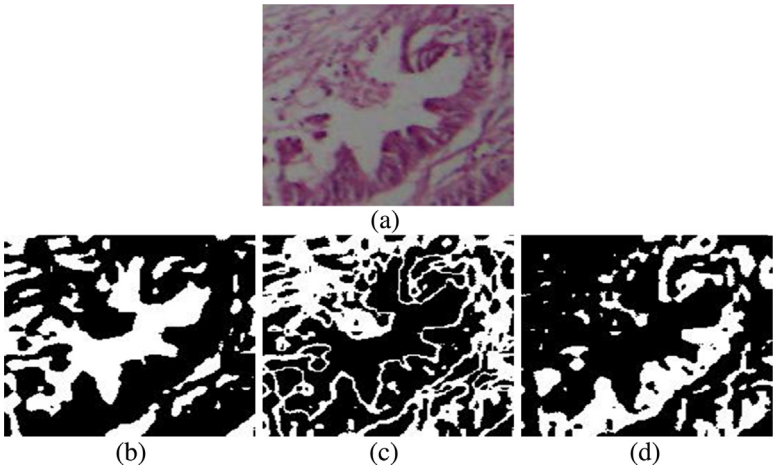

Fig. 3 – Image clustering using *K*-Means clustering algorithm: (a) original colon biopsy image, (b) white, (c) pink, and (d) purple clusters. (For interpretation of the references to color in the text, the reader is referred to the web version of the article.)

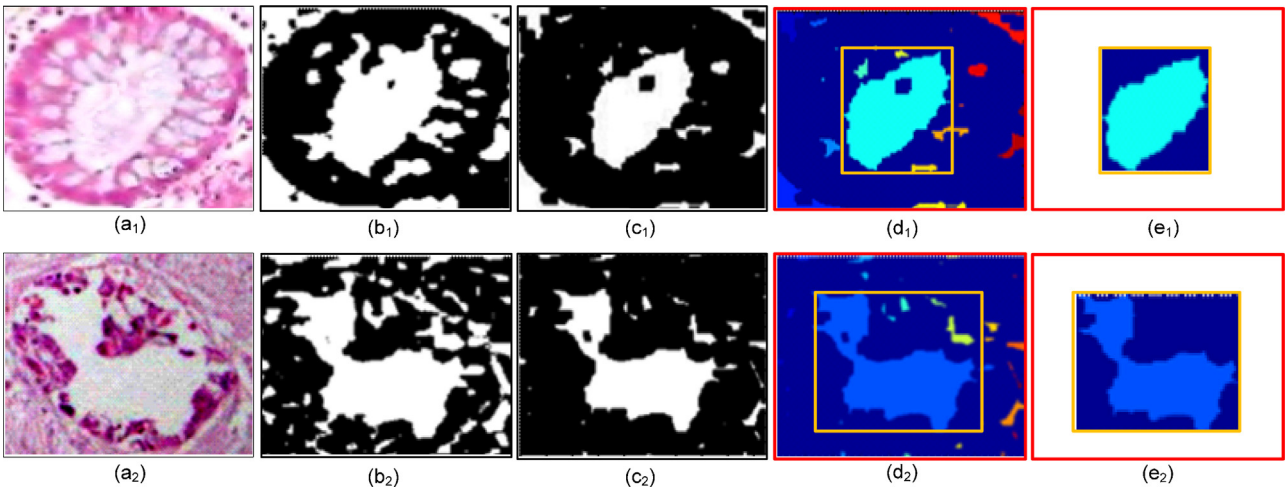

Fig. 4 – The process of lumen extraction from colon biopsy images: (a_1 and a_2) normal nad malignant colon biopsy images, respectively, (b_1 and b_2) white clsuters, (c_1 and c_2) eroded clusters obtained after erosion using a disk shaped structuring element, (d_1 and d_2) generated connected components, the largest connected component is in the rectangular region, (e_1 and e_2) final lumen obtained after hole filling in the largest connected component.

3.2.1. *Lumen inner concavity (LIC) and lumen outer convexity (LOC)*

The lumen, shown in Fig. 1(a), is the central part of a colon tissue, and is surrounded by various elliptic shaped epithelial cells. The lumen has near-elliptic shape in normal colon tissue, whereas lumen and epithelial cells merge with each other in well-differentiable cancer grade of colon tissues, by this means generating larger white patches. Therefore, we have proposed two novel structural descriptors, which quantify the deviation of malignant lumen from the underlying elliptic shape in normal tissue. The process of the extraction of these descriptors comprises the following three main steps.

1. *Lumen extraction*

   Lumen is the largest structure in white cluster of colon tissue (see Fig. 1(a)), therefore, it is extracted from the white cluster only. The lumen extraction process for sample normal and malignant colon biopsy images is shown in the first and second row of Fig. 4, respectively. Fig. 4(a_1) and (a_2) shows original normal and malignant colon biopsy images, respectively. Fig. 4(b_1) and (b_2) shows white clusters of these images. As shown for the cluster of normal sample in Fig. 4(b_1), lumen is connected with epithelial cells, therefore, morphological erosion is performed in order to disjoint them. The erosion is performed with a disk-shaped

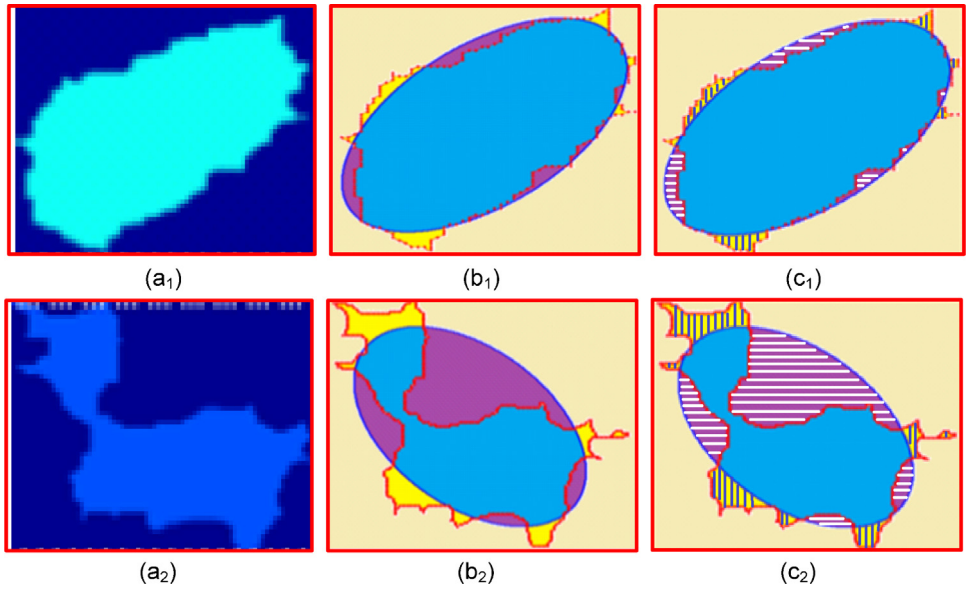

Fig. 5 – The process of least square ellipse fitting to the extracted lumnes: (a_1 and a_2) lumen of normal and malignant colon biopsy images, respectively, (b_1 and b_2) least-square ellipses fitted to the extracted lumens, (c_1 and c_2) quantification of lumen shape.

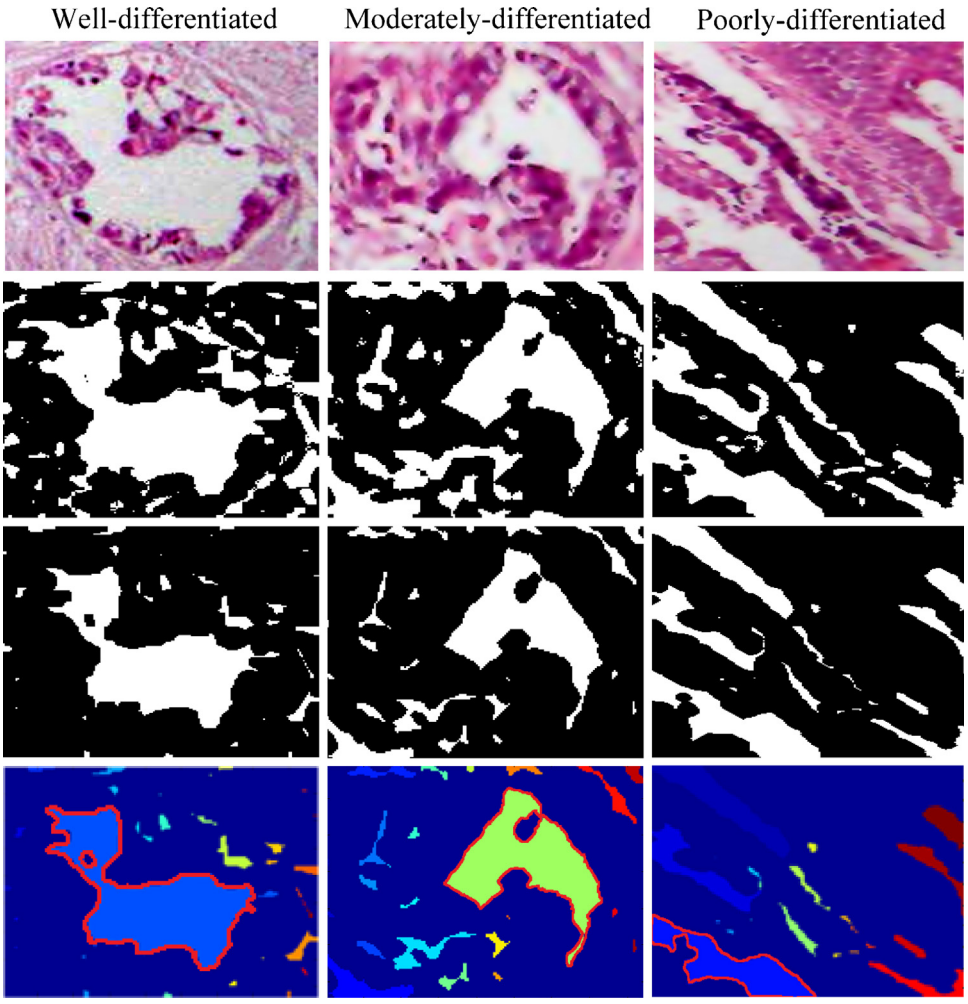

Fig. 6 – The process of computing LPWC and LPWI features from colon biopsy images of various cancer grades: 1st column: well-differentiable colon biopsy image, 2nd column: moderate-differentiable colon biopsy image, 3rd column: poor-differentiable colon biopsy image, 1st row: original colon biopsy images, 2nd row: white clusters of given colon biopsy images, 3rd row: eroded clusters obtained after erosion of white clusters, 4th row: different connected components in the images; the largest component is shown with an encircled boundary.

structuring element, and is formulated in the following equation.

$$C^* = C \odot s \tag{1}$$

where $\odot$ denotes the erosion operator, C represents the white cluster, s is the disk shaped structuring element of radius 2, and C^* is the resultant cluster. The result of erosion is shown in Fig. 4(c_1) and (c_2). In the next step, lumen is extracted by finding the largest connected component in C^*. Fig. 4(d_1) and (d_2) shows all the connected components for the sample images. The largest connected component is segregated from other components of the image. The largest connected components for the samples images have been shown in rectangular boundary in Fig. 4(d_1) and (d_2). Finally, the lumen is obtained by applying the hole filling process to the extracted largest component. The final lumen is shown in Fig. 4(e_1) and (e_2) for the sample images.

2. *Ellipse fitting*
   In this step, a least square ellipse is fitted to the contour of the extracted lumen in order to capture the underlying geometry of the extracted lumen. The least square method fits such a candidate ellipse to the extracted contour that minimizes the pixel-wise distance between the contour and the candidate ellipse in the least square sense. The least square ellipse fitting process is graphically represented in Fig. 5 for the lumens shown in Fig. 4 (e_1) and (e_2), respectively.
3. *Quantification of lumen shape*
   The fitted ellipse will approximate the contour of extracted lumen better for normal colon biopsy images. Whereas, the irregular contour of the malignant lumen will deviate considerably from the fitted ellipse. This can be easily observed from the ellipses fitted to sample malignant and normal images shown in Fig. 5(b_1) and (b_2), respectively. In this step, such deviation is quantified in terms of the following two novel structural descriptors.

- *Lumen inner concavity* (*LIC*): LIC measures the number of pixels of the extracted lumen, which are not part of the fitted ellipse. These pixels are shown by the area filled with vertical lines in Fig. 5(c_1) and (c_2). LIC descriptor is computed as follows:

$$LIC = \frac{(U - \bar{\bar{E}}) \cap L}{\bar{\bar{E}}} \tag{2}$$

where E, L and U represent the sets of pixels in fitted ellipse, the extracted lumen, and the bounding box around extracted lumen, respectively. The terms $(U - \bar{\bar{E}}) \cap L$ and $\bar{\bar{E}}$ represent the cardinality of sets $(U - E) \cap L$ and E, respectively. The factor $(U - \bar{\bar{E}}) \cap L$ is divided by $\bar{\bar{E}}$ in order to normalize the LIC measure with respect to the size of fitted ellipse.

- *Lumen outer convexity* (*LOC*): LOC measures the number of pixels of the fitted ellipse, which are not part of the extracted lumen. These pixels are shown by the area filled with horizontal lines in Fig. 5(c_1) and (c_2). LOC descriptor is computed as follows.

$$LOC = \frac{(U - \bar{\bar{L}}) \cap E}{\bar{\bar{E}}} \tag{3}$$

where E, L and U are the same as used in LIC feature. The terms $(U - \bar{\bar{L}}) \cap E$ and $\bar{\bar{E}}$ represents the cardinality of sets $(U - L) \cap E$ and E, respectively. The factor $(U - \bar{\bar{L}}) \cap E$ is divided by $\bar{\bar{E}}$ in order to normalize the LOC measure with respect to the size of fitted ellipse.

3.2.2. *Lumen circularity (LUC)*

The near-elliptic shape of lumen in a normal colon tissue can be considered as a slightly squeezed circle, therefore, the measure of lumen circularity can also be exploited as a valuable structural descriptor for discerning normal and malignant tissues. This measure can exploit the structural variation between normal and malignant lumen from a different perspective compared to the proposed LIC and LOC descriptors. LUC is a measure of the circularity ratio of the extracted lumen, and it can be calculated using the following mathematical expression.

$$LUC = \frac{4 \times \Pi \times a}{p^2} \tag{4}$$

where p and a represent the perimeter and area of the extracted lumen, respectively. Fig. 4(a_1) and (a_2) presents sample normal and malignant colon biopsy images, and Fig. 4(e_1) and (e_2) shows the corresponding extracted lumens. The extracted lumen of the normal image has LUC of 0.87, which is higher compared to the LUC (0.26) of the malignant image.

3.3. *Cancer grading phase*

The purpose of cancer grading phase is to identify well, moderate, and poor grades of malignant colon biopsy images. In this phase, two novel area based structural descriptors are calculated from white cluster of colon biopsy images. The descriptor values are then formulated, and images are classified into different cancer grades based on these feature values. These structural descriptors are described in detail in the following text.

3.3.1. *Lumen area based descriptors*

The distribution of all the cytological tissue constituents such as epithelial cells, non-epithelial cells, and connecting tissues of a colon biopsy image changes due to malignancy. However, the variation that lumen and epithelial cells experience is significant among all. This variation is also notable among different grades of progressing colon cancer. In the well-differentiable grade, epithelial cells and lumen merge together, in this manner resulting in larger white patches. With the increase in severity of colon cancer, the white patch splits into multiple smaller patches, which are dispersed among other constituents of colon tissue. Therefore, we intend to measure the variation between different colon cancer grades by calculating the area spanned by the largest patch in the white cluster of colon biopsy image.

For calculating area based features, largest connected component is extracted from white cluster using the same method as already discussed in Section 3.2.1 for extraction of lumen. This process is graphically illustrated in first, second and third columns of Fig. 6 for sample colon biopsy images with well-, moderate-, and poor-differentiable grades, respectively. The first row in the figure shows original colon biopsy images, whereas second row shows white clusters corresponding to these images. Similar to the cancer detection phase, erosion is applied on the clusters in order to disjoint epithelial cells from lumen. The eroded clusters are shown in the third row of Fig. 6. Finally, the largest connected component is obtained by converting eroded cluster into connected components. The fourth row in Fig. 6 shows all the components and the corresponding largest connected component (encircled by boundary) for well, moderate, and poor grade samples.

Later, the expansion in lumen and epithelial cells has been quantized in terms of two novel structural descriptors. Let I, W and C represent the sets of pixels in the image, white cluster, and the extracted largest component, respectively, the structural descriptors are defined as under.

- *Lumen percentage with cluster* (*LPWC*): LPWC is the percentage area occupied by the largest white connected component in the white cluster. By area, we mean number of occupied pixels.

$$LPWC = \frac{\bar{\bar{C}}}{\bar{\bar{W}}} \times 100 \tag{5}$$

where $\bar{\bar{C}}$ and $\bar{\bar{W}}$ represent the cardinality of set C and W, respectively.

- *Lumen percentage with image* (*LPWI*): LPWI is the percentage area occupied by the largest white connected component

in the whole image. By area, we mean number of occupied pixels.

$$LPWI = \frac{\bar{\bar{C}}}{\bar{\bar{I}}} \times 100 \quad (6)$$

where $\bar{\bar{C}}$ and $\bar{\bar{I}}$ represent the cardinality of set C and I, respectively.

4. Performance measures

The classification capability of the various features proposed in this work has been quantitatively evaluated using various performance measures such as accuracy, sensitivity, specificity, Mathew's correlation coefficient (MCC), F-score, and receiver operating characteristics curve (ROC). Normal and malignant images correspond to negative and positive samples, respectively. Therefore, true positive (TP) and true negative (TN), respectively, are the number of correctly classified malignant and normal images. Similarly, false positive (FP) and false negative (FN), respectively, represent the number of incorrectly classified normal and malignant images.

Accuracy: The classification accuracy is a measure of usefulness of a technique [17]. It depends upon the number of correctly classified samples, and is calculated using the following equation.

$$Accuracy = \frac{TP + TN}{N} \times 100$$

where N is the total number of colon biopsy images.

Sensitivity: Sensitivity is a measure of the ability of a technique to correctly identify positive samples [17]. It can be calculated using the following equation.

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

The value of sensitivity ranges between 0 and 1, where 0 and 1mean worst and best recognition of positive samples, respectively.

Specificity: Specificity is a measure of the ability of a technique to correctly identify negative samples [17]. It can be calculated using the following equation.

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

The value of specificity ranges between 0 and 1, where 0 and 1mean worst and best recognition of negative samples, respectively.

Matthews correlation coefficient (MCC): MCC is a measure of the efficacy of binary class classifications [18]. It can be calculated using the following mathematical expression.

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

The valid range of MCC is between −1 and +1, where +1, −1 and 0, respectively, correspond to best, worst, and random prediction.

F-Score: F-score is a weighted average of precision and recall [17]. It is a measure of the correctness of classification algorithm. F-score can be calculated using the following mathematical expressions.

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

$$F\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The valid values of F-score range between 0 and 1, where 0 and 1, respectively, mean worst and best classification.

ROC: An ROC curve is a standard way for graphical representation of the classification performance of a system [17]. It has been used in various research studies in the past for analyzing the performance of classifiers [19,20]. It characterizes the system over its entire operating range, and is created by plotting true positive rate (TPR) against false positive rate (FPR). In practical applications, the ROC curve provides a degree of freedom to select the operating point which best accomplishes the requirements of the application.

5. Results and discussion

This section presents the performance of the proposed structural descriptors for detection and grading of colon cancer on the dataset presented in Section 5.1. The performance has been evaluated in terms of various well-known measures such as accuracy, sensitivity, specificity, Matthews correlation coefficient, F-score and receiver operating characteristics (ROC) curve. Section 5.2 is about experimental setup, and enlights the training/testing data formulation, and the adopted classification methodology. The optimal values of the parameters involved in the descriptors' computation and the classification stages of the proposed CCD system have been empirically selected. The selection process is discussed in detail in Section 5.3. The results of cancer detection and cancer grading phases have been described in Sections 5.4 and 5.5, respectively. The computational time requirements of various phases of the proposed CCD system have been calculated through experimentation, and have been discussed in Section 5.6. In order to validate the effectiveness of the proposed features, they have been compared with various state-of-the-art features. These results are summarized in Section 5.7. Similarly, performance of the selected SVM classifier has been compared with various other classifiers, and is described in Section 5.8. In Section 5.9, the performance of the proposed CCD system has been described in comparison with state of the art cancer detection and grading techniques. The computations have been carried out on Intel Core i7 with 3.4 GHz processor and 12 GB RAM. Matlab computational software for 64-bit windows has been used in all the experiments.

5.1. Dataset

The biopsy samples used in this research work have been collected from the pathology division of Rawalpindi Medical College (RMC), Rawalpindi, Pakistan in the years 2010–2012. The available 68 permanent colon biopsy sections for the above mentioned time period have been collected without any discrimination of gender, race, and age. The thickness of tissues in the biopsy slide is 5–6 μm, and the samples have been

Table 1 – Statistics for the dataset.

Parameters	Values
Number of images	174
Distribution of images	92 malignant
	82 normal
Grades of malignant images	23 poor-differentiable
	44 moderate-differentiable
	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

stained with Hematoxylin and Eosin. The imaging equipment has been provided by PAEC General Hospital, Islamabad. The magnification factor of the objective lens of the microscope has been set to 10×, and RGB images of colon have been captured at 600 × 800 resolutions. A dataset of 174 variable size microscopic RGB images has been extracted from the images captured at 600 × 800. The dataset has been prepared under the guidance of classified histopathologist (Imtiaz Ahmad Qureshi, Assistant Professor) of the RMC college. The ground truth labels have been assigned to the given dataset by three histopathologists, namely, Dr. Imtiaz Ahmad (RMC), Dr. Rahat Abbas (RMC), Brig. Shoaib Nayyar Hashmi (Armed Forces Institute of Pathology, Rawalpindi). Out of total of 174 images, the histopathologists have perfect agreement for 170 images (80 normal and 90 malignant). Similarly, for the grading dataset, the histopathologists have perfect agreement for 85 images. For the remaining images, labels have been assigned based on the decision of majority of histopathologists. Furthermore, Kappa statistic has also been calculated in order to determine the variability in the diagnosis of different histopathologists. The Kappa value for cancer detection and grading datasets is 0.9692 and 0.9194, respectively. Table 1 provides detailed statistics of the dataset.

5.2. *Experimental setup*

Features are extracted from the given images, and are scaled in the range 0–1 for better classification results. Next training/testing data is formulated using the cross-validation methodology, and is classified into respective classes. The adopted training/testing and classification methodology are explained in the following text.

5.2.1. *Training/testing data formulation*

The formulation of training and testing dataset is an important phase. Generally, three cross-validation methods, namely sub sampling, independent dataset test, and jackknife are used by the statisticians to predict the efficacy of a classifier in practical applications. Chou [21] has demonstrated in a recent review that among the three methods, jackknife cross-validation is supposed to be least subjective and rigorous because of its ability to yield a distinct result for a given dataset. In medical diagnosis systems, it is extremely desirable to achieve unique output for a sample no matter how many times the sample is being tested. Therefore, jackknife test has been increasingly used by the researchers for estimating the goodness of various classifiers in medical diagnostic systems [22–24]. Thus, the jackknife 10-fold cross-validation has been used both in the cancer detection and cancer grading stages of the proposed CCD system to examine the probable success rates of the classifiers.

5.2.2. *Classification model*

SVM is a well-known classifier that has been widely used in the past for classification of medical images [25,26]. SVM is based on the principle of finding a decision surface that has maximum distance to the closest points of the two classes in the training data set [2]. The training objective of SVM is to find such an optimal decision surface such that the classification error for new test samples is minimized.

In the proposed CCD system, RBF kernel of SVM has been employed as classifier both in the cancer detection and cancer grading stages. RBF is a local kernel of SVM, and its performance is based on Euclidean distance between the training samples. Let x_i and x_j be the training samples of the dataset, the RBF kernel is defined as follows:

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$$

RBF kernels has two adjustable parameters (γ and c). The parameter 'c' represents the cost of the constraint violation associated with the data points occurring on the wrong side of decision surface, drawn by RBF SVM. The parameter γ shows the width of the Gaussian function.

5.3. *Selection of optimal values for system parameters*

The performance of the proposed CCD system depends on several parameters. Therefore, it is always preferred to find optimal values of these parameters for best performance. In the following subsections, the selection of optimal values of these parameters is presented in detail.

5.3.1. *Selection of parameters for structural descriptors*

The morphological operation of erosion is performed while computing descriptors from colon biopsy images both in the cancer detection and grading phases. A disk shaped structuring element is used for erosion, the radius of which (say α) plays a vital role in the separation of epithelial cells and lumen. Therefore, it is essential to find optimal radius of the structuring element. Fig. 7(a) and (b) shows the effect of varying α on classification accuracy of cancer detection and grading phases, respectively. It can be seen from Fig. 7(a) that the classification performance is maximum for $\alpha = 2$. Similarly, Fig. 7(b) demonstrates that the structural descriptors used in the grading phase are not sensitive to the size of structuring element after $\alpha = 2$, where the system attains highest classification accuracy. Therefore, a disk-shaped structuring element with $\alpha = 2$ has been used both in the cancer detection and grading phases of the proposed CCD system.

5.3.2. *Selection of parameters for classification*

The RBF kernel employs two critical parameters, namely c (cost of constraint violation), and γ (width of Gaussian function), which should be tuned prior to classification for optimal performance. In this research work, grid search method [27] has been employed for this purpose, and the optimal values

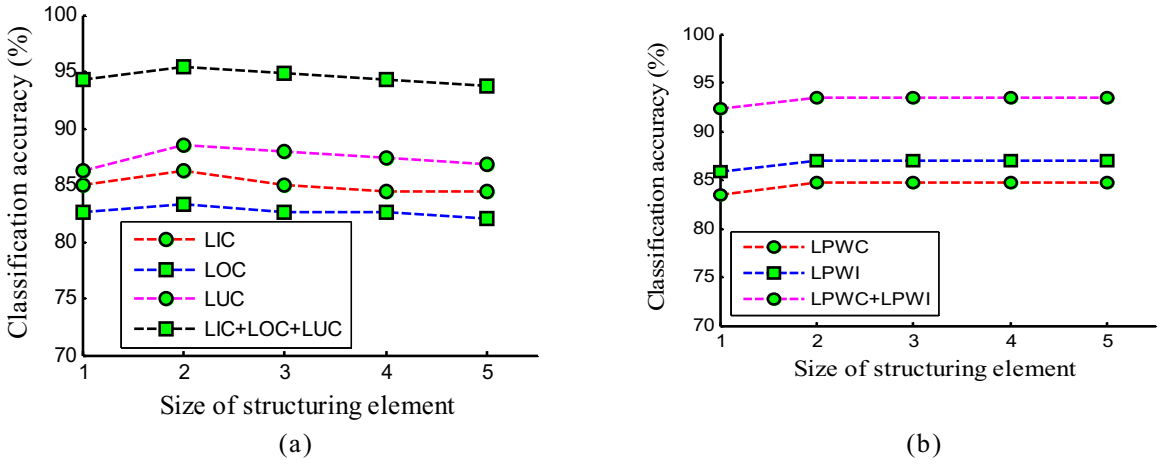

Fig. 7 – Classification performance of the proposed individual and hybrid descriptors as a function of α: (a) cancer detection and (b) grading phases.

Table 2 – Classification results of the descriptors proposed for cancer detection.

	Accuracy	Sensitivity	Specificity	MCC	*F*-score
LIC	86.21	0.880	0.841	0.723	0.871
LOC	83.33	0.837	0.829	0.666	0.842
LUC	88.51	0.902	0.866	0.769	0.892
LIC + LOC + LUC	95.40	0.956	0.951	0.908	0.957

of c and γ parameters have been obtained by adjusting the grid range of $c = [1, 2, \ldots, 100]$ with $\Delta c = 1$, and $\gamma = [0.001, 0.003, \ldots, 0.099]$ with $\Delta\gamma = 0.002$.

5.4. *Performance analysis of the cancer detection phase*

In this section, we present our results of the descriptors proposed for cancer detection on the dataset presented in Section 5.1. The descriptors have been used individually as well as in the hybrid form (i.e. LIC + LOC + LUC). The results of the descriptors in terms of various performance measures (see Section 4) are reported in Table 2. The results demonstrate that the proposed individual descriptors i.e. LIC, LOC, and LUC have good discerning capability for classification of images into normal and malignant categories. The classification success rate of 86.21%, 83.33%, and 88.51% has been observed for LIC, LOC, and LUC descriptors, respectively. When the structural descriptors are hybridized, they all reinforce each other, by this way boosting the sample recognition rate up to 95.40%. Both normal and malignant samples are identified equally well, as verified by almost equal values of sensitivity (0.956) and specificity (0.951) achieved using hybrid descriptor. The results of MCC (0.908) and *F*-score (0.957) are also in accordance with the classification accuracy. The high classification performance measures prove that the proposed descriptors truly capture the variation in the morphology of normal and malignant colon tissues.

In addition to the results given in Table 2, the performance of the proposed descriptors has also been analyzed in terms of ROC curves, which are shown in Fig. 8. The ROC curves are well above the diagonal line, thereby proving the worthiness of the proposed descriptors for detection of colon cancer. Furthermore, the ROC curves demonstrate that the hybrid descriptor (LIC + LOC + LUC) has better discrimination power as its ROC curve is well above that of its individual counterparts. Therefore, the results for cancer detection have been reported only for the hybrid descriptor in subsequent text.

Fig. 9 shows a few normal and malignant images, which are correctly identified by cancer detection phase of the proposed CCD system. The normal images shown in Fig. 9 have standard geometry; the lumen can be identified separately from other constituents, and is near-elliptic in shape. Therefore, small pixel-wise difference exists between the extracted lumen and the fitted ellipse, thereby resulting in smaller values of the proposed LIC and LOC descriptors. Secondly, the lumen seems quite compact in these images, by this means resulting in larger values of LUC descriptor for these images. On the contrary, lumen and epithelial cells have been merged with each other in the malignant images shown in Fig. 9, and

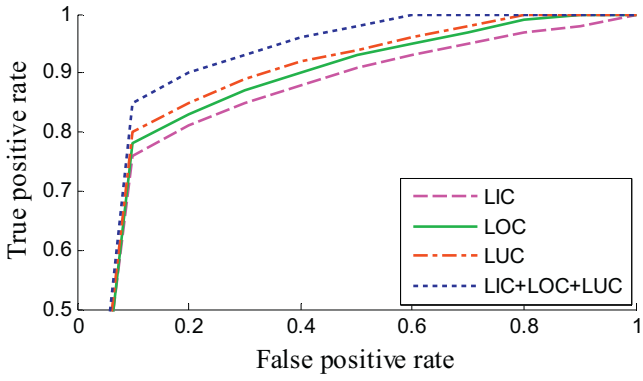

Fig. 8 – Classification performance of the descriptors proposed for cancer detection in terms of ROC curve.

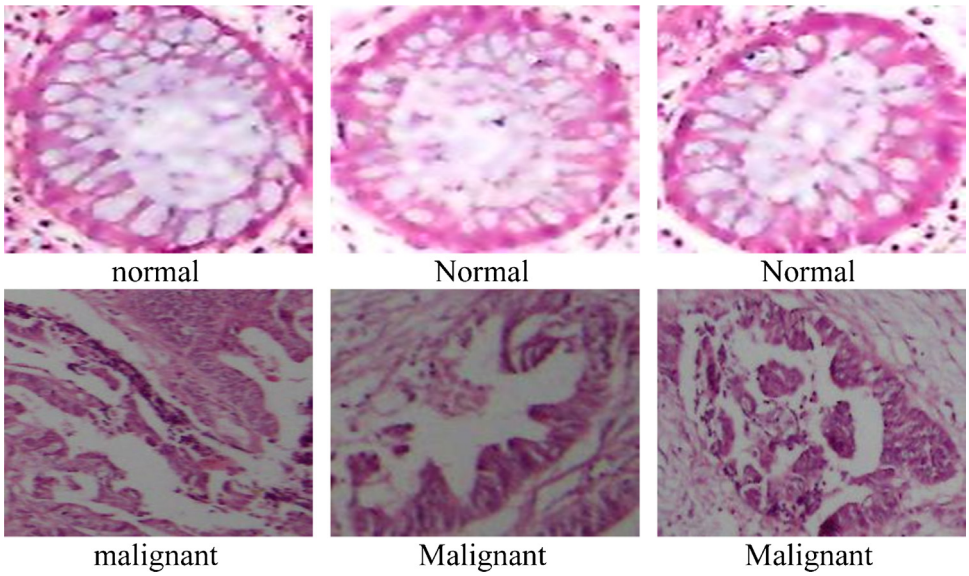

Fig. 9 – Normal and malignant images, which are correctly identified by the proposed cancer detection phase using the hybrid feature vector (LIC + LOC + LUC).

the elliptic shape of lumen has diminished, which results in larger values of LIC and LOC descriptors, and smaller value of LUC descriptor. This shows that the variation in the shape of lumen between normal and malignant images is quite handily exploited by the proposed descriptors. Further, malignant images of various caner grades shown in Fig. 9 prove that the proposed descriptors are robust against various cancer grades, and quite correctly identify malignant images regardless of the severity of cancer grade.

Fig. 10 presents a few images incorrectly classified by the cancer detection phase of the proposed CCD system. The three columns of Fig. 10 show colon biopsy images, the extracted lumens, and the fitted ellipses, respectively. Fig. 10(a_1) and (a_2) shows normal colon biopsy images, whereas Fig. 10(a_3) shows a malignant image. The close observation reveals that lumen in the normal images is not in the near-elliptic shape. These images lie at the boundary between well-differentiable malignant and the normal state of colon. These images deviate from the standard geometry of normal colon tissues, therefore, the system is deficient in accurately identifying them. Similarly, the malignant image in Fig. 10(a_3) is misclassified by the proposed technique, as the shape of the lumen is irregular, but still near-elliptic in this image.

5.5. *Performance analysis of the cancer grading phase*

In this section, we present our findings about the performance of the proposed LPWC and LPWI descriptors for classification of malignant images into well-, moderate- and poor-differentiable cancer grades. The performance of the proposed descriptors has not only been measured for overall malignant dataset, but also for individual cancer grades. Both individual and hybrid descriptors (i.e. LPWC + LPWI) have been investigated for cancer grading as already done in the cancer detection phase. The confusion matrices for LPWC, LPWI, and LPWC + LPWI are given in Table 3. The results in Table 3 demonstrate that the proposed descriptors have excellently captured the biological variation that lumen undergoes with progressing cancer. The diagonal entries in the confusion matrices verify that the proposed descriptors have correctly classified a larger subset of well-, moderate- and poor-differentiable images.

Table 3 – Confusion matrices for LPWC, LPWI and LPWC + LPWI descriptors.

	Well	Moderate	Poor
Confusion matrix for LPWC			
Well	20	2	1
Moderate	3	36	5
Poor	1	2	22
Confusion matrix for LPWI			
Well	21	1	1
Moderate	2	38	4
Poor	1	3	21
Confusion matrix for LPWC + LPWI			
Well	22	1	0
Moderate	2	39	3
Poor	0	2	23

Table 4 reports classification accuracy of the proposed descriptors for overall malignant dataset and for individual cancer grades. The recognition rate of 93.47% for overall dataset, and the recognition rates of 95.65%, 88.64% and 92.00%, respectively, for well-, moderate-, and poor-differentiable cancer grades show that the proposed descriptors are quite effective at discriminating various colon

Table 4 – Classification accuracy of the descriptors proposed for cancer grading.

	Well	Moderate	Poor	Overall
LPWC	86.96	81.82	88.00	84.78
LPWI	91.30	86.36	84.00	86.95
LPWC + LPWI	95.65	88.64	92.00	93.47

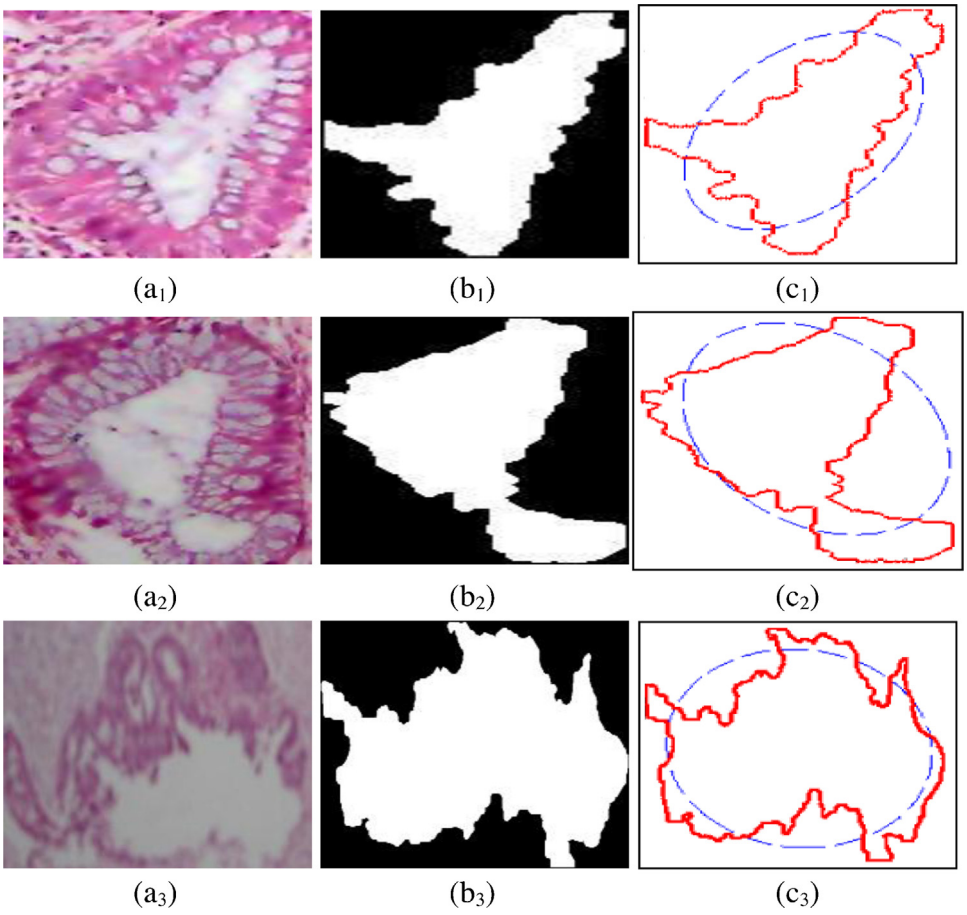

Fig. 10 – Colon biopsy images incorrectly classified by the proposed cancer detection phase: (a_1–a_3) original colon biopsy images, (b_1–b_3) extracted lumens, and (c_1–c_3) least-square ellipses fitted to the extracted lumens.

cancer grades. Further, the results demonstrate that the starting grade (well-differentiable) is easy to distinguish compared to other grades of colon cancer as shown by the classification success rate of 95.65%. Thus, the proposed CCD system can be effective in detection of colon cancer in early stage, thereby improving the disease prognosis.

Table 5 presents the performance of the proposed descriptors for detection of individual cancer grades in terms of a few other performance measures. As concluded from the results of Table 4, the well-differentiable grade is identified most conveniently by the proposed CCD system. Further, the better values of sensitivity, specificity, and F-score for well-, moderate and poor-differentiable cancer grades show the efficacy of the proposed descriptors for discerning various colon cancer grades.

Table 5 – Classification results of the descriptors proposed for cancer grading in terms of sensitivity, specificity, and F-score.

	Sensitivity	Specificity	F-score
LPWC			
Well	0.8696	0.9420	0.8511
Moderate	0.8182	0.9167	0.8571
Poor	0.8800	0.9104	0.8302
LPWI			
Well	0.9130	0.9565	0.8936
Moderate	0.8636	0.9167	0.8837
Poor	0.8400	0.9254	0.8235
LPWC + LPWI			
Well	0.9565	0.9710	0.9362
Moderate	0.8864	0.9375	0.9070
Poor	0.9200	0.9552	0.9020

Finally, the performance of the proposed descriptors has also been analyzed in terms of the ROC curve. In this context, the ROC curves of individual and hybrid descriptors are shown in Fig. 11. Large area under the ROC curves demonstrates that the proposed descriptors have good discrimination power. Further, the ROC curve of hybrid descriptor is well above the ROC curve of other descriptors. Thus, it can be concluded from the results presented in Tables 3–5 and the ROC curves shown in Fig. 11 that the proposed LPWC and LPWI descriptors are good at capturing variations amongst different cancer grades, but this distinguishing capability is further enhanced when the descriptors are combined to form a hybrid descriptor. Therefore, the results for cancer grading have been reported only for the hybrid descriptor (LPWC + LPWI) in subsequent text.

Fig. 12 presents a few sample colon biopsy images, which are correctly graded by the proposed CCD system. It can be seen from the figure that the well-differentiable image contains large white patch (lumen), thereby resulting in larger

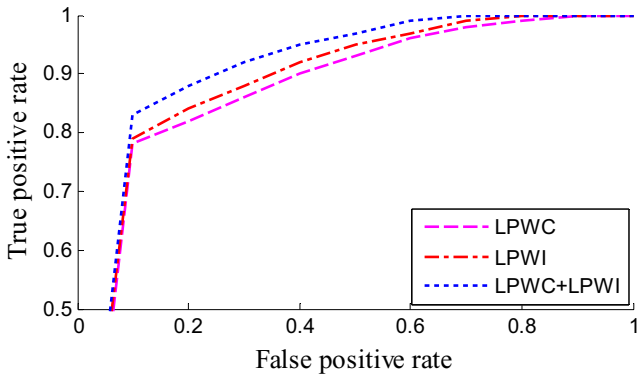

Fig. 11 – Classification performance of the descriptors proposed for cancer grading in terms of ROC curve.

values of the proposed LPWC and LPWI descriptors. Whereas, the lumen becomes smaller and dispersed as the cancer grade progresses to moderate- and poor-differentiated states as shown in Fig. 12(b) and (c), respectively. This characteristic is effectively exploited by the proposed hybrid descriptor (LPWC + LPWI) in order to successfully identify the grade of these images.

Fig. 13 contains a few malignant images, which are incorrectly classified by the cancer grading phase. The images in the figure contain two labels each; first one is the true grade of the image, whereas second label (pointed by arrow) refers to the grade predicted by the proposed system. It can be seen that the images deviate from the normally observed behavior of progressing grades, i.e. the largest white region is very small for well-differentiable image and vice versa for poor-differentiable image. Therefore, the proposed system incorrectly identifies the grades of these images.

5.6. *CPU time requirements of the proposed CCD system*

In this section, the computational time requirements of the proposed CCD system have been investigated. In this context, the CPU time elapsed from the start to the end of each phase of the system is measured in seconds. Generally speaking, the proposed system first transforms an image into a set of clusters in the preprocessing phase, and operates only on white cluster throughout the rest of the phases, therefore, computational time will be less. For the pre-processing phase, the CPU time involved in applying *K*-Means on each image is measured separately, and is shown in Fig. 14. The average time consumed in pre-processing of an image is 0.29 s, which reveals that the computational time for the pre-processing stage is nominal.

In the cancer detection and grading phases, the time consumed during the extraction of corresponding structural descriptors, and the classification time taken by RBF classifier for each descriptor type have been separately measured in seconds. The results are provided in Table 6. The time involved in computing descriptors has been separately calculated for each image, and average and standard deviation has been calculated for the whole dataset. The extraction time for hybrid descriptor is merely the sum of extraction time of constituent descriptors. The results in Table 6 reveal that the average computational time required for the extraction of hybrid descriptor of cancer detection and grading phases is 0.1182 ± 0.1335 and 0.0111 ± 0.0056s, respectively. The extraction time is quite nominal, which proves that the proposed descriptors are computationally tractable. The classification

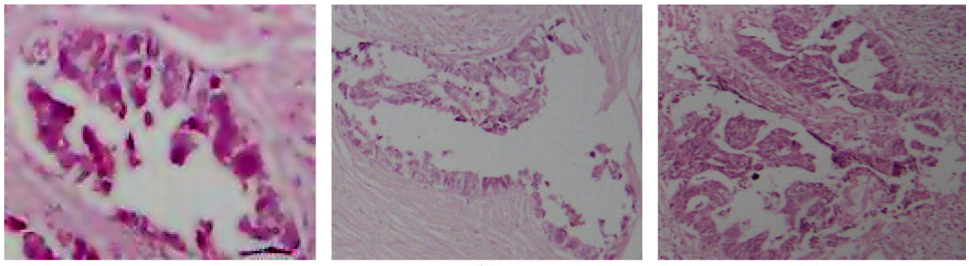

Fig. 12 – (a) Well-, (b) moderate-, and (c) poor-differentiable images, which are correctly classified by the proposed cancer grading phase.

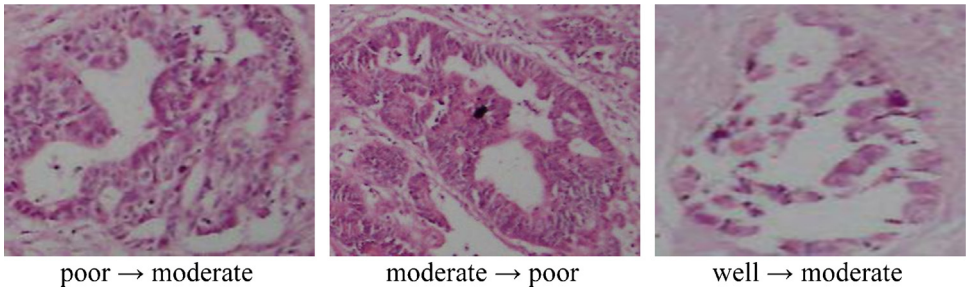

Fig. 13 – Examples of a few malignant images, which are incorrectly classified by the proposed caner grading phase.

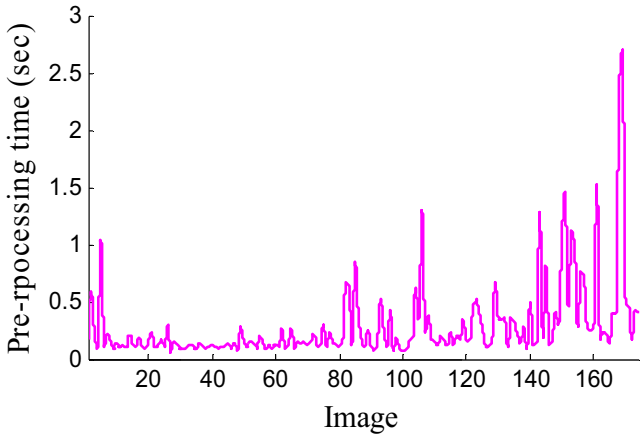

Fig. 14 – The CPU time involved in pre-processing (K-Means clustering) of colon biopsy images.

time has been separately calculated for each descriptor of the cancer detection and grading phases, and is in fact the time of applying cross-validation on complete descriptor set for computing labels of samples using RBF classifier.

5.7. *Performance comparison with existing feature types*

In order to validate the effectiveness of the proposed features for colon cancer detection and grading, the classification results using the proposed features have been compared with several state-of-the-art feature types. These features include morphological, texture, statistical moments, scale invariant feature transform (SIFT), elliptic Fourier descriptors (EFDs) and histogram of oriented gradients (HOG) features. First four statistical moments i.e. mean, standard deviation, skewness and kurtosis have been used for comparison. The texture features used for comparison include features of entropy, correlation, contrast, and uniformity, and morphological features include features of area, perimeter, eccentricity, Euler number, convex area, compactness, orientation, length of major and minor axes.

Several parameters are involved in the computation of these features such as number of SIFT points for SIFT features, size of gray-level co-occurrence matrix for texture features, harmonics levels for the EFD features, etc. So, to have to fair comparison of the proposed features with these state-of-the-art methods, optimal values of parameters involved in these features have been used for calculating these features. Table 7 summarizes the results for existing features and the proposed features. The results show that the proposed features, which capture the underlying geometry of colon tissues, have yielded better results compared to the features, which capture general texture from an image such as texture, morphological, and statistical moments. Nonetheless HOG features have shown good results for cancer detection, but have not been able to produce better results for cancer grading.

Table 6 – CPU time (s) required for the extraction and classification of descriptors proposed for cancer detection and grading.

Structural descriptors	Feature extraction time	Classification time
Cancer detection phase		
LIC	0.0363 ± 0.0237	42.367894
LOC	0.0369 ± 0.0243	68.236677
LUC	0.0451 ± 0.1082	36.196408
LIC + LOC + LUC	0.1182 ± 0.1335	35.231589
Cancer grading phase		
LPWC	0.0057 ± 0.0029	18.840314
LPWI	0.0055 ± 0.0027	15.707479
LPWC + LPWI	0.0111 ± 0.0056	19.786876

Table 7 – Comparison of the proposed features with existing feature types.

	Cancer detection	Cancer grading
EFDs	85.06	82.60
SIFT	82.76	85.86
HOG	94.86	86.95
Texture	88.45	84.78
Morphological	83.50	83.69
Statistical moments	82.39	82.60
Proposed features	95.40	93.47

5.8. *Performance of the proposed features by employing some other classifiers*

This section investigates the classification capability of some other classifiers for colon cancer detection and grading. In this context, various classifiers such as K-nearest neighbor (KNN), probabilistic neural network (PNN), decision tree, Bayesian, neural network, and other kernels of SVM such as linear, polynomial, and sigmoid have been used both for cancer detection and grading. The cancer detection and grading capability using these classifiers has been shown in Table 8. The results in the table show that RBF kernel of SVM has much better results compared to other classifiers, which verifies that SVM is the best choice of classifier for cancer detection and grading in the proposed methodology.

5.9. *Performance comparison of the proposed CCD system with existing techniques*

In this section, we have performed a performance comparison of the proposed CCD system with previously published

Table 8 – Classification accuracy of the descriptors proposed for cancer detection and grading using various classifiers.

	Cancer detection	Cancer grading
KNN	78.26	73.26
PNN	75.86	73.03
RBF SVM	95.40	93.47
Linear SVM	84.48	81.52
Sigmoid SVM	85.05	82.60
Polynomial SVM	88.50	85.86
Bayesian	86.78	83.69
Neural network	87.35	88.04
Decision tree	89.08	89.13

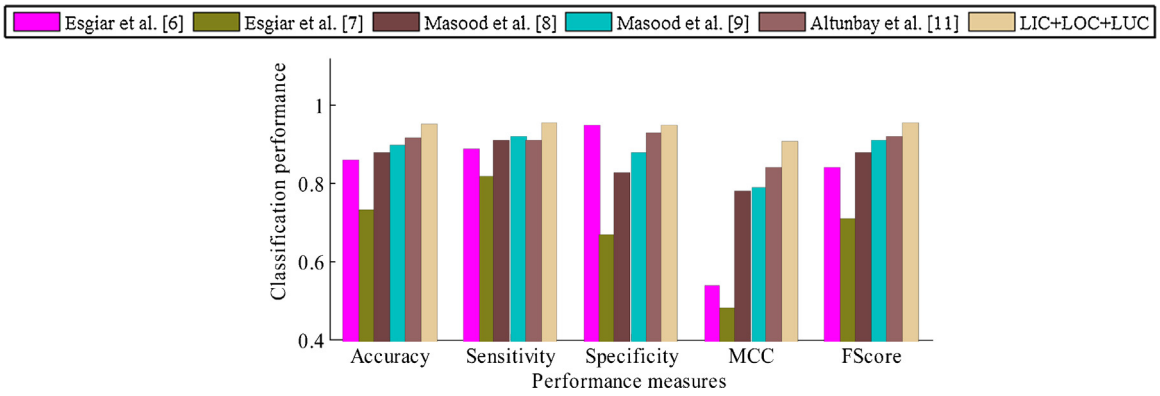

Fig. 15 – Performance comparison of the cancer detection phase of the proposed CCD system with existing cancer detection techniques.

approaches of colon cancer detection and grading. To this end, we have selected five existing colon cancer detection techniques [6–10] for comparison with the cancer detection phase of the proposed CCD system. The selected techniques have been implemented in Matlab computational software, and their performance has been evaluated on the dataset described in Section 5.1. In order to develop a fair comparison with the proposed technique, optimal values of parameters involved in selected methodologies have been obtained by exploring a suitable range of values for these parameters. Fig. 15 provides a performance comparison of the cancer detection phase of the proposed CCD system with previous cancer detection techniques. The classification accuracy of the proposed cancer detection technique is 98.28%, which is higher compared to the best accuracy (91.28%) reported by previous techniques. The better cancer detection capability of the CCD system in comparison with previous techniques may be attributed to the proposed descriptors, which quite handily exploit the structural variation between normal and malignant tissues. The proposed LIC and LOC descriptors capture the deviation of lumen from the standard elliptic shape, therefore, are better at identifying normal and malignant tissues. Similarly, LUC descriptor that measures the compactness of lumen is able to quantify the variation between the shape of normal and malignant lumen.

The performance of the cancer grading phase of the proposed CCD system has also been compared with existing cancer grading techniques [10,12]. In these techniques, malignant colon biopsy images have been divided into two grades of colon cancer (low grade and high grade) by using linear SVM. Therefore, in order to establish a comparison with the proposed CCD technique that caters three cancer grades, we have computed the same set of descriptors proposed in these works, but have performed multiclass classification using the RBF classifier. The comparison of these techniques with the proposed CCD system is graphically presented in Fig. 16.

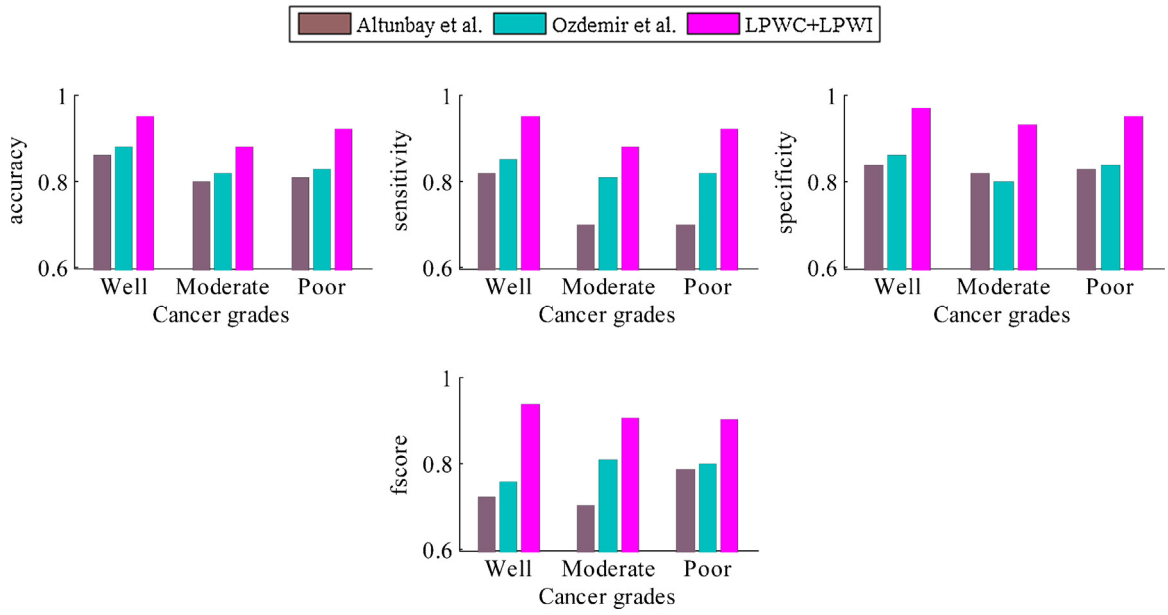

Fig. 16 – Performance comparison of the cancer grading phase of the proposed CCD system with existing cancer grading technique.

Compared with the performance of previous techniques, the proposed CCD system has shown better results in terms of all the performance measures as demonstrated in Fig. 16. The previous techniques make graph of all the cytological tissue components and valuate some descriptors on the graph as already discussed in the introduction section. The comparison in Fig. 16 shows that the descriptors defined on the shape and other properties of lumen are more effective and robust than those defined on graph of all the cytological tissue components. The lumen based features are not only computationally tractable, but also improve classification performance compared to the features proposed in previous techniques.

6. Conclusion

This research article presents a computer-aided colon cancer diagnostic system for automatic detection and grading of colon cancer. It proposes to incorporate the background knowledge about the morphology of normal and malignant tissues into the classification process. To this end, a new set of structural descriptors, which quantify the location, shape and area of lumen in the colon biopsy images, have been introduced. The proposed CCD system has been validated on 92 malignant and 82 normal images. The experimental results show that the proposed system has good cancer detection and grading capability with the best possible accuracies of 95.40% and 93.47%, respectively. The results in terms of various other performance measures also verify the worthiness of proposed descriptors for detection and grading of colon cancer. The hybrid structural descriptor has also been employed, and the results show that the rich combination of individual descriptors in a hybrid descriptor significantly improves the system performance. Compared with several contemporary techniques, the results show that the proposed CCD system is more effective in detection and grading of colon cancer. There are several possible future directions. First possibility is to classify colon biopsy images acquired at multiple magnification factors of the microscopic lens. Second is to acquire more biopsy slides and test the proposed technique on a larger dataset. Third and the most important possibility is to do gland level segmentation by incorporating our previously proposed method of colon biopsy image segmentation [28], and the method proposed by Fu et al. [29], and later doing classification based on features of multiple glands.

REFERENCES

[1] Cancer Facts and Figures, American Cancer Society, 2012.
[2] Colon cancer risk factors, in: C.C. Alliance.
[3] 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.
[4] 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.
[5] 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.
[6] 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. Inf. Technol. Biomed. 2 (1998) 197–203.
[7] A.N. Esgiar, R.N.G. Naguib, B.S. Sahrif, M.K. Bennett, Fractal analysis in the detection of colonic cancer images, IEEE Trans. Inf. Technol. Biomed. 6 (2002) 54–58.
[8] K. Masood, N. Rajpoot, H. Qureshi, K. Rajpoot, Co-occurrence and morphological analysis for colon tissue biopsy classification, in: Proc. 4th International Workshop on Frontiers of Information Technology, Islamabad, Pakistan, 2006, pp. 211–216.
[9] K. Masood, N. Rajpoot, Texture based classification of hyperspectral colon biopsy samples using CLBP, in: Proc. International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 2009, pp. 1011–1014.
[10] 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.
[11] A.B. Tosun, M. Kandemir, C. Sokmensuer, C.G. Demir, Object-oriented texture analysis for the unsupervised segmentation of biopsy images, J. Pattern Recogn. 42 (2009) 1104–1112.
[12] E. Ozdemir, C.G. Demir, A hybrid classification model for digital pathology using structural and statistical pattern recognition, IEEE Trans. Med. Imaging 32 (2013) 474–483.
[13] Duda, Image Processing.
[14] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 790–799.
[15] D. Comaniciu, P. Meer, Robust analysis of feature spaces: color image segmentation, in: International Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997, pp. 750–755.
[16] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: International Conference on Computer Vision and Pattern Recognition, Hilton Head, SC, USA, 2000, pp. 142–149.
[17] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, London, 2005.
[18] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phase lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.
[19] S. Rathore, M.A. Iftikhar, M. Hussain, A. Jalil, Classification of colon biopsy images based on novel structural features, in: 9th International Conference on Emerging Technologies, Islamabad, Pakistan, 2013.
[20] S. Rathore, M.A. Iftikhar, M. Hussain, A. Jalil, A novel approach for ensemble clustering of colon biopsy images, in: 11th International Conference on Frontiers of Information Technology, Islamabad, Pakistan, 2013.
[21] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.
[22] S.S. Sahu, G. Panda, A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction, Comput. Biol. Chem. 34 (2010) 320–327.
[23] L. Nanni, A. Lumini, D. Gupta, A. Garg, Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (2012) 467–475.
[24] M. Hassan, A. Chaudhary, A. Khan, M.A. Iftikhar, Information gain based fuzzy C-means clustering and classification of carotid artery ultrasound images, Comput. Methods Progr. Med. 113 (2013) 593–609.

[25] M. Tahir, A. Khan, H. Kaya, Protein subcellular localization in human and hamster cell lines: employing local ternary patterns of fluorescence microscopy images, J. Theor. Biol. 340 (2014) 85–95.
[26] M. Hayat, A. Khan, Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition, J. Theor. Biol. 271 (2011) 10–17.
[27] 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.
[28] S. Rathore, M. Hussain, A. Khan, A novel approach for colon biopsy image segmentation, in: IEEE (Ed.), Complex Medical Engineering, Beijing, China, 2013.
[29] H. Fu, G. Qiu, J. Shu, M. Ilyas, A novel polar space random field model for the detection of glandular structures, IEEE Trans. Med. Imaging 33 (2014) 764–776.