A Recent Survey on Colon Cancer Detection Techniques

Saima Rathore, Mutawarra Hussain, Ahmad Ali, and Asifullah Khan

Abstract—Colon cancer causes deaths of about half a million people every year. Common method of its detection is histopathological tissue analysis, which, though leads to vital diagnosis, is significantly correlated to the tiredness, experience, and workload of the pathologist. Researchers have been working since decades to get rid of manual inspection, and to develop trustworthy systems for detecting colon cancer. Several techniques, based on spectral/spatial analysis of colon biopsy images, and serum and gene analysis of colon samples, have been proposed in this regard. Due to rapid evolution of colon cancer detection techniques, a latest review of recent research in this field is highly desirable. The aim of this paper is to discuss various colon cancer detection techniques. In this survey, we categorize the techniques on the basis of the adopted methodology and underlying data set, and provide detailed description of techniques in each category. Additionally, this study provides an extensive comparison of various colon cancer detection categories, and of multiple techniques within each category. Further, most of the techniques have been evaluated on similar data set to provide a fair performance comparison. Analysis reveals that neither of the techniques is perfect; however, research community is progressively inching toward the finest possible solution.

Index Terms—Colon biopsy, colon cancer, texture, hyperspectral, gene, blood serum analysis

1 Introduction

Large intestine performs wide variety of functions, ranging from breakage of large molecules to nutrients and water absorption [1], [2]. Colon is one major constituent of large intestine, and its cancer is a major reason of deaths in western and industrialized world [3]. There are many reasons of colon cancer, like, chain smoking, increasing age such as age above 50 years, family history of colon cancer, low intake of fruits, and heavy intake of red meat and fats [4], [5].

Traditionally, colon cancer is diagnosed using microscopic analysis of histopathological colon samples. In such an examination, pathologists observe the colon samples under microscope to detect malignancy, and assign cancer grade depending upon the level of organizational changes they observe in tissues. But, the manual examination has a few limitations. First, it is subjective because quantitative measures such as cancer grades/stages mainly depend on the visual assessment of pathologists. Second, it has inter/intra observer variation in grading [6], [7], [8], [9]. Such vulnerabilities in the manual process result in need of automatic colon cancer diagnostic techniques [10], [11], which could provide second opinion to the pathologists in the short run and could serve as an independent trustworthy system for detection of cancer in the long run.

Automatic detection of colon cancer has two major directions: segmentation and classification. In segmentation, heterogeneous colon samples are segregated into homogenous regions based on spatial distribution of tissues in the images. Next, normal and malignant labels are assigned to the regions based on certain features. In the literature, several approaches exist for medical image segmentation such as pixel based, region based, and graph based. Pixel-based methods divide image pixels into different clusters based upon their colors using various approaches like watershed transform [12], clustering [13], [14], adaptive segmentation [15], and thresholding [16]. Region-based segmentation methodologies use similar approach, but they maintain connectivity between pixels of similar clusters. Well-known techniques of this category include splitting and merging [17], and region growing [18]. Graph-based techniques [19], [20] assume image pixels as nodes of a graph, and weight between them as similarity between pixels. Segmentation then involves graph partitioning into subgraphs while minimizing cost functions. In classification, colon samples are divided into normal and malignant categories based upon certain features. Classification and segmentation may be followed by cancer grading step, in which quantitative cancer grades are assigned to the samples depending upon certain quantitative measures.

Automated diagnostic systems have been proposed for cancer detection in various body parts such as brain [21], [22], [23], [24], [25], breast [26], [27], [28], [29], [30], cervical [31], prostate [32], [33], and lungs [34]. In this connection, several techniques have also been proposed for colon cancer detection. A subset of these techniques, known as texture-analysis-based techniques, have exploited the noteworthy

- *S. Rathore is with the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan, and the Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir. E-mail: saimarathore_2k6@yahoo.com.*
- *M. Hussain, A. Ali, and A. Khan are with the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan. E-mail: {mutawarra, asif}@pieas.edu.pk, ahmadali1655@hotmail.com.*

Manuscript received 7 Mar. 2013; revised 26 June 2013; accepted 4 July 2013; published online 22 July 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2013-03-0071. Digital Object Identifier no. 10.1109/TCBB.2013.84.

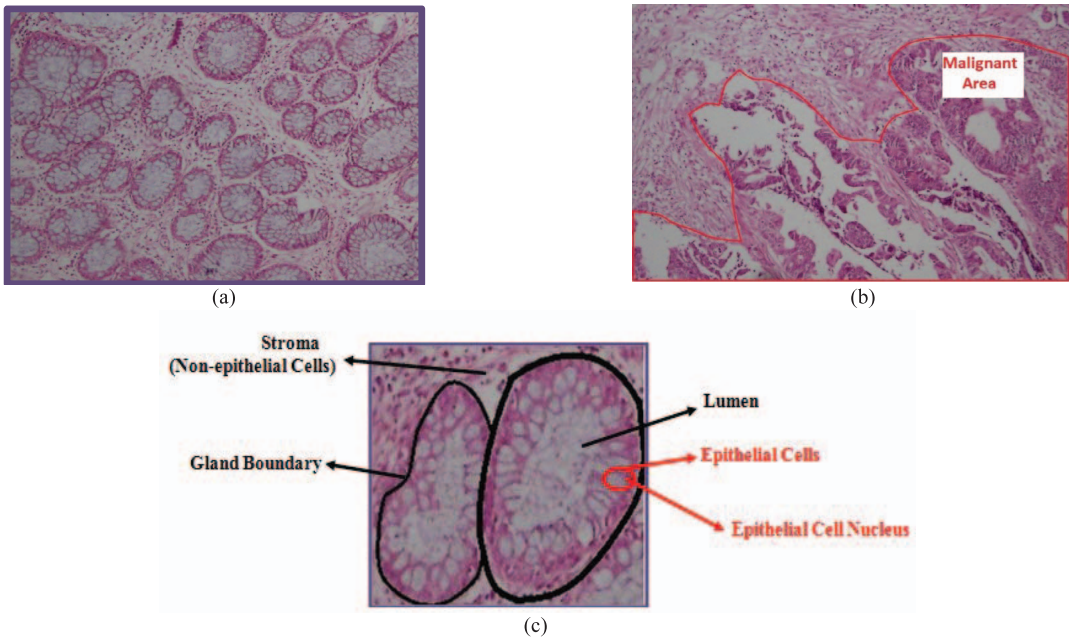


Fig. 1. (a) Normal and (b) malignant colon tissues, and (c) constituents of normal colon tissue.

contrast between texture of normal and malignant parts of colon biopsy images for cancer detection [35], [36], [37], [38], [39], [40], [41]. Several statistical and morphological features, which have proven to be good discriminators in the past [42], [43], have been investigated in these studies. There are a few other techniques, called object-oriented (OO) texture analysis-based techniques, which incorporate background knowledge of normal and malignant tissue organization into the diagnostic process [44], [45], [46], [47], [48], [49], [50], [51], [52]. Focus of a few studies is to perform hyperspectral analysis on colon biopsy images, and classify samples into different classes [53], [54], [55], [56], [57], [58], [59]. Another interesting and quite promising method of colon cancer detection is simultaneous analysis of thousands of human genes to detect genetic alterations responsible for cancer [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73]. Some researchers have also exploited the variation in chemical composition and resulting Raman spectra of normal and malignant serum for cancer diagnosis [74], [75], [76], [77], [78], [79], [80], [81], [82], [83]. Another method of colon cancer detection is to simulate the neural activity happening in the brain by analyzing organization of colon biopsy images [84], [85].

In 2009, Demir and Yener [86] reported a survey in which a few issues involved in automated colon cancer diagnosis based on histopathological images have been discussed. However, this paper does not cover colon cancer detection techniques in detail. Further, extensive research has been carried out in the field of automated colon cancer diagnosis in the last two decades, but so far, a comprehensive survey in this field has not been reported. Therefore, a latest review of the research in the field of colon cancer diagnosis is highly desirable. This work, thus, fulfills the basic need of researchers working in this field by providing an extensive discussion on classical as well as contemporary techniques. Hence, this paper serves a needy purpose in this connection.

In this paper, we divide colon cancer detection techniques into five major categories. In each category, detailed explanation of most of the techniques along with a healthy discussion on their merits and demerits is provided. Further, performance comparison of different categories and of several techniques within each category is presented in detail. The novel contribution of this study is the implementation and evaluation of colon cancer detection techniques on same data set.

Remainder of this paper is organized as follows: Section 2 highlights organization of colon tissues. Section 3 presents a detailed insight into existing colon cancer detection techniques. Section 4 provides a performance review of the techniques. Section 5 evaluates the performance of various colon cancer detection techniques on prepared data sets, and Section 6 concludes the paper.

2 Benign and Malignant Tissue Organization

Normal colon tissues have well-defined organizational structure. However, this arrangement varies in case of cancer. Variation usually depends upon the cancer stage. Initial cancer stages deform the cells very little. Therefore, are harder to detect. On the other hand, advanced stages significantly deform the cells, thereby making their detection easier.

Fig. 1 presents normal and malignant colon tissues. Deformation introduced by cancer is clearly visible in Fig. 1b, but tissues are organized in normal colon sample (see Fig. 1a).

Fig. 1c presents three constituents (epithelial cells, nonepithelial cells, and lumen) of normal colon tissue. Epithelial cells usually surround lumen and form glandular structure, whereas nonepithelial cells, called stroma, lie in between these structures. There are five stages of colon cancer(0,I-IV) according to National Cancer Institute [87]. These stages are similar to the Duke's stages (0,A-D) [88],

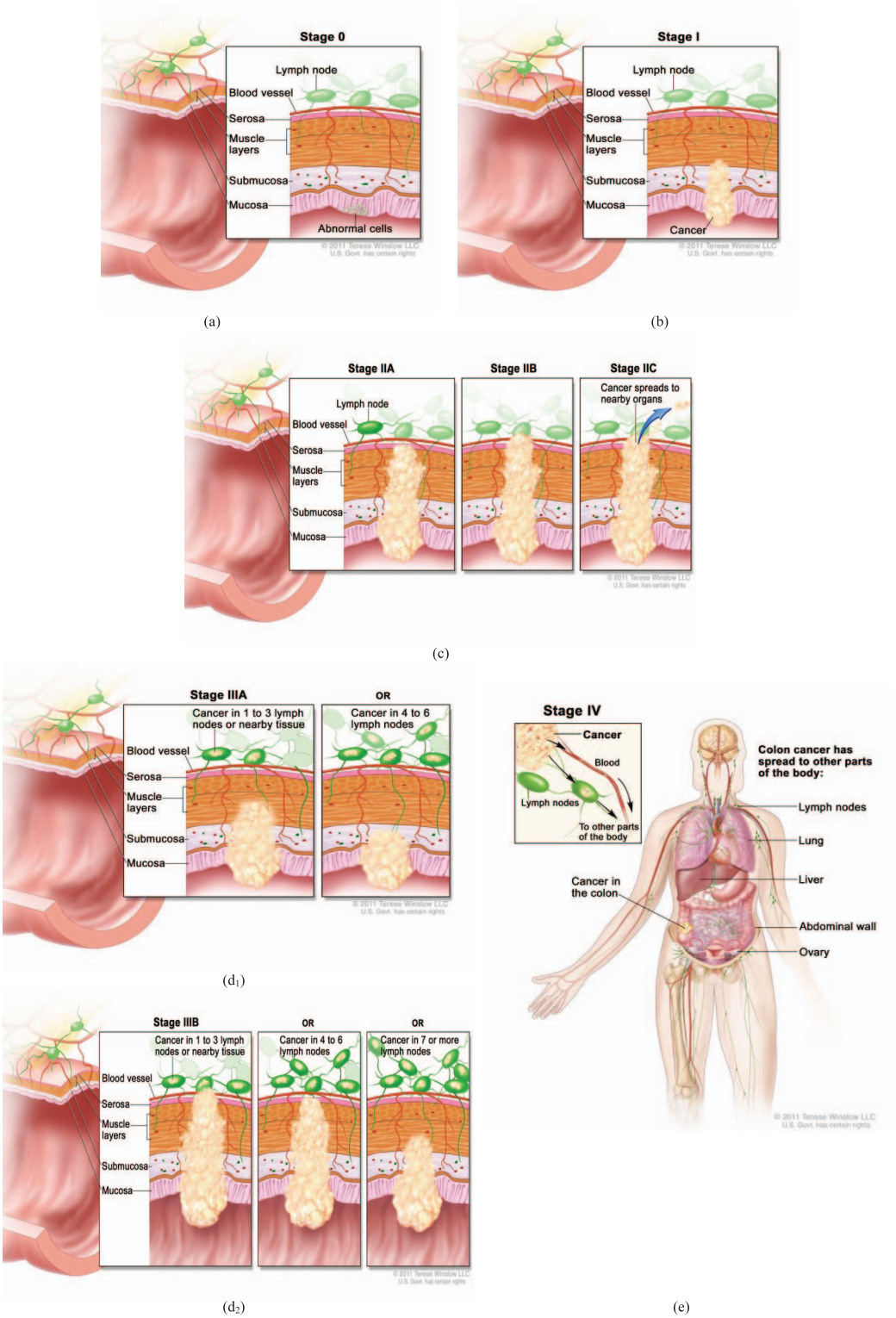


Fig. 2. Different stages of colon cancer: (a) Stage 0, (b) Stage I, (c) Stage II, (d_1, d_2) Stage III, and (e) Stage IV.

[89]. Stage 0 is the earliest stage in which cancer just starts to develop. It is still restricted to the innermost lining of colon. In Stage I, cancer has reached to the middle layer of colon. In Stage II, cancer has reached beyond the middle layer. Cancer is called Stage III if it reaches lymph nodes, and is found in at least three of them. Stage IV is the final stage, wherein cancer has reached other body parts such as lungs and liver. Fig. 2 demonstrates these stages.

3 Colon Cancer Segmentation and Classification Techniques

Generally, there are five major categories of colon cancer detection techniques depending upon the underlying data set and adopted methodology. These categories include spectral analysis, texture analysis, gene analysis, serum analysis, and OO texture analysis. Texture, hyperspectral,

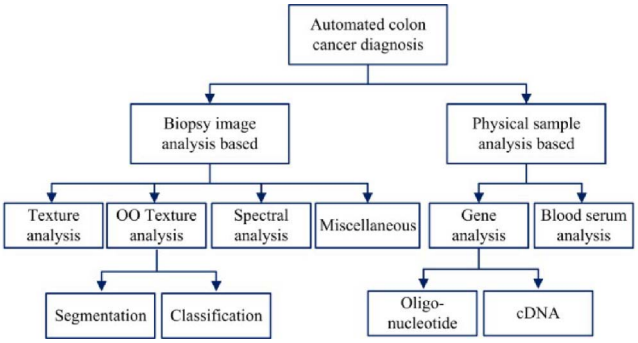


Fig. 3. Top-level breakdown of colon cancer classification/segmentation techniques.

and OO texture analysis-based techniques work on images. Therefore, these techniques have been referred as image analysis-based techniques in this survey. A few techniques, which work on colon biopsy images but are not well established, have been discussed in the miscellaneous category. Gene and serum analysis-based techniques analyze physical sample for cancer detection, therefore, have been named physical sample-based colon cancer detection techniques. A broad level categorization of these techniques is presented in Fig. 3, and the following text explains these techniques in detail.

3.1 Texture Analysis-Based Techniques

Texture is a combination of repeated patterns with regular/irregular frequency [90]. There is a significant variation in the texture of normal and malignant colon tissues. A few researchers have exploited this variation in their respective studies, and achieved good classification on colon biopsy data sets. Entropy and correlation have been commonly used to quantize texture. However, several other parameters are also in practice. Detailed information on textural features may be found in the classical book of digital image processing [91]. Well-known texture analysis-based colon cancer detection techniques have been summarized in the following text.

Esgiar et al. [35] proposed a promising method of colon cancer detection by using textural features. In their work, original colon biopsy images of size 512×512 are further divided into four subimages of size 256×256, and the subimages having little tissue content are excluded. Gray-level cooccurrence matrix (GLCM) is then calculated for each subimage. Normalized GLCM is used to determine textural features of angular second moment, contrast, correlation, inverse difference moment, dissimilarity, and entropy by using (1)-(6), respectively,

$$\sum_{i=1}^{R}\sum_{j=1}^{R} P_{ij}^2, \tag{1}$$

$$\sum_{i=1}^{R}\sum_{j=1}^{R} (i-j)^2 P_{ij}^2, \tag{2}$$

$$\sum_{i=1}^{R}\sum_{j=1}^{R} \frac{(i-\mu_r)(j-\mu_c)P_{ij}}{\sigma_r \sigma_c}, \tag{3}$$

$$\sum_{i=1}^{R}\sum_{j=1}^{R} \frac{P_{ij}}{[1+(i-j)^2]}, \tag{4}$$

$$\sum_{i=1}^{R}\sum_{j=1}^{R} |i-j| P_{ij}, \tag{5}$$

$$-\sum_{i=1}^{R}\sum_{j=1}^{R} P_{ij} log P_{ij}, \tag{6}$$

where P_{ij} is the ijth entry in the normalized GLCM matrix. i and j are the integer pixel numbers along its rows and columns, which are quantized up to level R. μ_r and μ_c are the mean of row sums and column sums, respectively. Likewise, σ_r and σ_c are the standard deviations of row sums and column sums, respectively. The reported classification accuracy is 90.2 percent for a combination of correlation and entropy by using linear discriminate analysis (LDA) classifier. Esgiar et al. [36] further extended their previous work, and employed geometric and texture features. Geometric features comprise features of shape and orientation. Texture features encompass energy, inertia and homogeneity, and are calculated from GLCM of the image. The reported classification accuracy for geometric and texture features is 80 and 90 percent, respectively.

Esgiar et al. improved their previous work [35] by adding image fractal dimensions to the feature set [37]. Concept of statistical scaling is used in calculating fractal dimensions, i.e., in a self-similar structure, a relationship "D_f" exists between scale factor of box size and number of boxes to which structure can be divided. Scale factor r is varied $(3, \ldots, 51)$, and the number of grid boxes *B(r)* containing the structure is counted. Relationship "D_f" is presented in the following equation:

$$D_f = \frac{\log(B(r))}{\log(1/r)}, \tag{7}$$

where D_f function is plotted for each image, and a regression line is fitted to the plotted points. Fractal dimension is calculated from the slope of the regression line. K-nearest neighbor (KNN) for $k = 2$, and LDA are used as classifiers with leave one out (LOO) approach of data formulation. Fractal analysis in combination with correlation and entropy improved accuracy up to 94.1 percent. An extended version of this study has been presented in another research paper [38].

Further, Kalkan et al. [39] combined texture and structural features to classify colon samples into normal, precancerous (adenomatous and inflamed) and malignant classes. In this work, 2,000 patches per class have been used. A total of 1,108 texture features are computed from each patch by evaluating 32 bins color channel histograms of R, G, B, H, S, V color components of raw image. Further, each patch is divided into 16 subpatches, and the following structural features are calculated per subpatch: the number of nuclei per tissue area, the individual and the pairwise ratio of each of stroma, cellular, and lumen to the tissue area. Further, forward feature selection strategy is applied to select meaningful features. Logistic regression classifier with equal class priorities is applied for classification, and 77.29, 82.25, 76.08, and 66.86 percent

classification accuracy is reported for adenomatous, malignant, inflamed, and normal classes.

Recently, Jiao et al. [40] proposed a simple and computationally tractable method of automatic colon cancer detection. In this work, colon biopsy images are initially converted to gray-scale images. Statistical features of mean and variance are computed form the images. Likewise, texture features of angular second moment, contrast, correlation, and entropy are computed from GLCM matrix in four different orientations of $\theta = 0^0, 45^0, 90^0, 135^0$. SVM with threefold cross-validation is employed to identify normal and malignant images. In testing phase, performance of each new feature is calculated by adding it to previous features. Results revealed that the composite feature vector comprising 18 features is more discriminative compared to individual features. Moreover, Ng et al. [41] carried out a research study with an aim to determine relationship between changes happening in the texture of malignant colon images and the survival rate of patients. In this work, texture features of standard deviation, entropy, uniformity, kurtosis, and skewness were extracted from pixel distribution histograms of contrast material-enhanced CT images. Kaplan-Meier analysis was performed to determine the relationship between these features and 5-year survival rate. The Cox proportional hazards model was used to assess independence of texture features from stage. This texture analysis process was followed up on the same cancer patients until their death. Analysis revealed that Kaplan-Meier survival plots for texture features are significantly different from each other. Results also revealed that features are independent from the cancer stage, and can be used to model malignancy in all the cancer stages. Further, results showed that fine texture features are associated with poorer 5-year overall survival rate for the patients of colon cancer.

3.2 OO Texture Analysis-Based Techniques

These techniques exploit background knowledge about size and spatial distribution of colon tissue components for segmentation and classification of colon biopsy images. These techniques have been further divided into segmentation and classification techniques.

3.2.1 *Object-Oriented Texture Analysis-Based Segmentation Techniques*

Initially, OO texture analysis was premeditated for segmentation of colon biopsy images. In this connection, object-oriented segmentation (OOSEG) [44] is the first OO texture analysis-based segmentation technique that comprises three well-defined phases, namely, object definition, texture definition, and segmentation. In object definition phase, k-means [92] is applied to divide image pixels into three clusters depending upon color intensities of tissue components such as purple-colored nuclei, white-colored lumen and epithelial cells, and pink-colored stroma. Circular primitives (example given in Fig. 4) are then located using a circle fitting process, summarized in Algorithm 1. Primitives of each cluster are further divided into two categories depending upon user-defined threshold, thereby resulting in six object types. In texture definition phase, two features called object size uniformity and object spatial distribution uniformity are calculated for each image pixel by considering six object types, thereby resulting in 12 features for each pixel. Segmentation algorithm initializes, grows, and finally merges seeds. In initialization phase, pixels having values of all the uniformity measures less than corresponding threshold (sum of mean and standard deviation) are marked as seeds. In seed growing, initial regions (seeds) are grown until they span the entire image. In region merging, two features, namely percentage area of each object type in a region, and percentage of combined areas of different objects that belong to particular cluster type in the region are calculated. Regions are merged if they are adjacent, and euclidean distance between their features is less than a merge threshold. Region merging phase merges small regions, and yields final segmentation results.

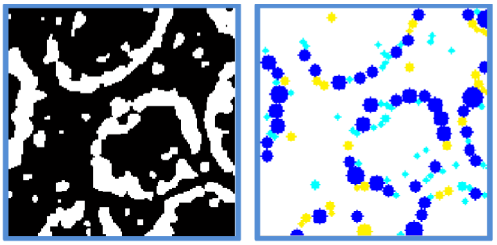


Fig. 4. (a) Purple cluster of a colon biopsy image, and (b) identified primitives.

Algorithm 1. Circle fitting algorithm.
For a given set of pixels P.

Step 1: Assign each particular pixel x_i to the largest possible circle that includes this particular pixel x_i and that is formed by only the pixels $xj \in$ **P.**

Step 2: Form connected components $\mathrm{C} = C_1, C_2, \ldots, C_N$ from the pixels such that the connected component *Ck* consists of the pixels that are assigned to the circle k.

Step 3: Eliminate the connected components smaller than an area threshold.

Step 4: For each component *Ck*, recursively call **Steps 1** and **2** considering only the pixels of this connected component (i.e., in **Step 1**, P will be a set of pixels belonging to the component *Ck*) until there is no change in the pixels of the component. (*There will be no change when a component is circular*.)

OOSEG [44] though produces reasonable results has a few limitations. For instance, it needs manual adjustment of parameters for each test image. Therefore, Tosun et al. proposed a new technique with an aim to identify a set of parameter values applicable to all image instances. In this work [45], object locating process is similar to OOSEG [44], but calculations in later stages are performed with reference to objects rather than pixels. In feature extraction phase, a set of twelve features as used in [44] is calculated for each object. Next, Voronoi diagram is constructed on the centroids of all the objects to determine initial seeds. Any two adjacent objects are grouped if euclidean distance between them is smaller than a predefined similarity threshold. Later, groups having number of objects larger than a threshold are declared seeds. Seeds are iteratively

grown until they cover the entire image. Final regions comprise objects. Therefore, Voronoi diagram of the objects is drawn to get pixel-based regions.

Demir et al. [46] proposed a valuable colon biopsy image segmentation technique. In this work, circle fitting process is similar to OOSEG [44] with one exception that only purple and white clusters are used to locate nucleus and lumen objects, respectively. Next, an object-graph [93] is constructed on these objects. Edges are assigned between each lumen object and its N closest lumen and N closest nucleus neighbors. For each lumen object L, features having information about areas, length of edges between L and its nucleus and lumen neighbors are extracted by considering neighbors within a circular window around L. These features are further used by the k-means algorithm to segregate lumen objects into "gland" and "nongland" classes. Objects of "gland" class are treated as initial seeds. Region growing process involves another object-graph that is constructed on nucleus objects. Starting from the initial lumen seeds, more lumen objects are added to the graph until an edge of the nucleus graph is encountered. Edges of the nucleus graph are used to stop region growing because glands are usually encircled by the nucleus objects, and encountering a nucleus object means that gland boundary is reached. In the end, false gland elimination, which has been proven to be effective in [94] and [95], is applied to remove false glands.

Tosun et al. [47] further improved OOSEG by employing graphs for quantifying spatial relationship between cytological tissue components. In the first step of this work, previously proposed methods of circle fitting [44], and graph generation [46], [93] are employed. In the second step, graph-edge runs are calculated. Graph-edge runs are based on the idea of gray-level run-length matrices [96]. Graph-edge run is a path that starts from an initial node, and contains all nodes reachable with a set of edges of the same type. For calculation of gray-level run-length matrix (GRLM), a circular window is hypothesized at center of a node, and then breadth first search is used to compute path for each particular edge type that lies within the window. In feature extraction phase, four features, namely, short-path emphasis (SPE), long-path emphasis (LPE), edge type nonuniformity (ETN), and path length nonuniformity (PLN) are computed. ETN and PLN help in determining the effect of edge type and path length distribution on texture, and occupy least values when the runs are uniformly distributed over all edge types and path lengths. Segmentation totally relies on objects instead of pixels, and comprises three steps: seed determination, region growing, and region merging. A window is centered on current object, and accumulated GRLM of the encircled object is calculated, which is used in feature calculation of current object. Initial seeds are determined by disconnecting pairs of adjacent objects having intermediate distance greater than a distance threshold, and removing components having lesser number of objects than a predefined threshold. Objects are merged to the seeds if they are adjacent, and euclidean distance between their features is smaller than merge threshold. Finally, Voronoi diagram of the objects is constructed to demarcate final region boundaries.

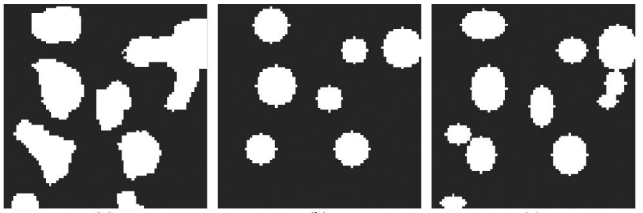


Fig. 5. (a) Cluster, objects located by using (b) circle fitting algorithm (943 pixels), and (c) ellipse fitting algorithm (1,359 pixels).

Simsek et al. [48] introduced cooccurrence features to quantify spatial relationship between objects in a colon biopsy image. Circular objects are located by using circle fitting algorithm [44]. A cooccurrence matrix is calculated for each object by placing a circular window on the object, and measuring the number of times objects of one type cooccur with objects of another type at a given distance d. Twenty-four cooccurrence features are extracted from the cooccurrence matrix. In this work, segmentation has been posed as a graph partitioning problem. Random objects are picked in different iterations to generate graphs. Segmentation is achieved by using these graphs. Finally, multiple results are combined to obtain final segmentation.

Recently, Rathore et al. [49] proposed a robust segmentation technique called modified object-oriented segmentation (MOSEG). In MOSEG, the underlying method of segmentation is similar to OOSEG [44] with a few enhancements. First, system parameters, which are set manually in OOSEG [44], are found separately for images captured at different magnification factors using genetic algorithm (GA). Second, epithelial cells, which were modeled as circles in previous techniques, are modeled as ellipses. Third, nearly elliptic shapes are also located in the images by using Algorithm 2 to cater blur. They validated their segmentation results at four different magnification factors. Algorithm 2 significantly increases the number of objects located in images compared to Algorithm 1, as presented in Fig. 5.

Algorithm 2. Ellipse fitting algorithm.

For a given set of pixels P for a given cluster L.

Step 1: Convert the set of pixels P into connected components $\mathrm{C} = C_1, C_2, \ldots, C_N$.
Eliminate the connected components smaller than an area threshold A.

Step 2: Find ellipses in the ith connected component C_i (**Step 2.A - Step 2.E**).
Step 2.A: Generate four patterns of simulated ellipses $EP_1(0^o), EP_2(45^o), EP_3(90^o), EP_4(135^o)$ starting with *CurrentMajorAxis* and *CurrentMinorAxis*.
(*CurrentMajorAxis* = *MaxMajorAxis* and *CurrentMinorAxis* = *MaxMinorAxis* in start).
Step 2.B: Locate elliptic primitives EP_1, EP_2, EP_3, EP_4 in the current connected component C_i one by one.
Step 2.C: Mark the pixels P' corresponding to elliptic primitives.
Step 2.D: Find remaining pixels $\mathrm{P}^* = \mathrm{P} - \mathrm{P}'$.
Step 2.E: Decrement axes values

TABLE 1
Morphological Features

Feature	Definition
Area	The number of pixels in a region.
Diameter	The largest distance between the boundary pixels of an object.
Extent	The ratio of pixels in the region to pixels in the total bounding box.
Orientation	The angle between the x-axis and the major axis of the ellipse.
Solidity	The ratio of the object area to the convex area.
Eccentricity	The ratio of the distance between the foci of the ellipse and its major axis length.
Euler number	The number of objects in the region minus number of holes in those objects.
Major axis	The length of the major axis of the ellipse.
Minor axis	The length of the minor axis of the ellipse.

(*CurrentMajorAxis* and *CurrentMinorAxis*) by 1 and continue (**Step 2.A**) finding ellipses in remaining pixels P^* until following criteria's meet:

1. Axes values reach minimum limit. (*MinMajorAxis* and *MinMinorAxis*).
2. No unassigned pixels are left, i.e., $P^* = []$.

Step 3: Continue (**Step 2**) finding ellipses in $(i+1)$th connected component C_{i+1}.

3.2.2 Object-Oriented Texture Analysis-Based Classification Techniques

Like segmentation, OO texture analysis-based techniques work equally well for classification of colon biopsy images. In this context, Altunbay et al. [50] proposed a novel texture features-based technique. In this work, previously proposed methods of circle fitting [44], and graph generation [46], [47], [48] are employed. A few structural features such as degree, average clustering coefficient (CC), and diameter are computed from the object-graph. Seven types of degrees are defined for each node. One degree type considers all edges, whereas remaining six degree types consider edges of particular colors. Averages of seven degrees for all the nodes constitute seven degree-based features for a single graph (image). CC is a measure of the connectivity in the neighborhood of a node. Four CCs are computed; first CC is computed by catering all nodes within the neighborhood, whereas remaining three are computed by considering nodes of unique colors. The clustering coefficient of a node n is defined as follows:

$$CC_n = \frac{2E_n}{d_n(d_n-1)}, \tag{8}$$

where d_n is the number of neighbors within the neighborhood, and E_n is the number of existing edges. It is noteworthy that E_n may be much lesser than d_n. Diameter is the longest of the shortest paths between any pair of nodes. Seven different diameters are calculated. The first diameter is computed by considering all edges, and the other six are computed by considering one particular edge type at a time. These 18 features are used to classify given samples by using linear SVM.

Recently, Ozdemir et al. [51] presented a resampling-based Markovian model for classification of colon biopsy images into normal, low grade and high grade cancer. In this work, perturbed samples (images) are generated from the original image. First-order discrete Markov model is employed to determine the posteriori probabilities of all the classes (normal, low grade, high grade) for a given perturbed sample. A class having highest posteriori probability is assigned to the perturbed sample. Finally, majority voting is employed to combine the classes of individual perturbed samples, and to determine the class of the original test sample. Moreover, Ozdemir and Demir. [52] presented another method of colon cancer detection based on object-graphs. Their idea is to make reference graphs [46], [47], [50] of a few images of normal glands, and then search query graphs of test images in the reference graphs. Query graphs are searched in the reference graphs by placing nucleus node of a query graph on each node of the reference graph. Three most similar graphs are found, and then based on the degree of similarity sample type is identified.

3.3 Hyperspectral Analysis-Based Techniques

Hyperspectral analysis-based techniques operate on selected spectral bands of colon biopsy images, and identify normal and malignant tissues. Hyperspectral data of colon biopsy images is collected by using hyperspectral imaging setup that consists of tuned light source [97].

In one of the earliest hyperspectral analysis-based techniques, the authors [53] segmented the hyperspectral colon biopsy images by using wavelet features, and provided quite promising visual results of segmentation. However, major focus of their research was the classification of hyperspectral colon biopsy images. In this connection, the authors [54] have accomplished classification of colon biopsy images using multiple kernels of SVM. Hyperspectral image cubes having size $1{,}024 \times 1{,}024 \times 20$ of colon tissues are acquired from hematoxylin and eosin (H&E) stained microarray. Overall process comprises four steps: preprocessing, segmentation, feature extraction, and classification. In preprocessing, FlexIA, a variant of independent component analysis, is applied for dimensionality reduction. In segmentation phase, extracted components are fed as input to the k-means clustering to produce $1{,}024 \times 1{,}024$ labeled images for each cube. In feature extraction, features given in Table 1 are extracted for 16×16 image patches, and for resolutions up to 256×256 to capture local as well as global details. Detailed

information about morphological features can be found in a classical book of image processing [91].

As a result, 4,096 features for each image cube and a total of 45,056 features for 11 selected cubes have been extracted. Thirty thousand features have been used for training and the remaining features have been used for testing. Gaussian, linear, and polynomial kernels of SVM have been used as classifiers. Results reveal clear superiority of Gaussian kernel with an accuracy of 87.5 percent.

In 2006, Masood et al. [55] have used GLCM and morphological features for colon tissue classification. Three distinct phases of their technique are segmentation, feature extraction, and classification. In segmentation phase, dimensionality of 3D cubes of hyperspectral image data is reduced using FlexICA. Imaging data is then divided into four clusters of nuclei, cytoplasm, glands, and stroma by using k-means. Two experiments are conducted. In the first experiment, morphological features of shape, size, orientation, and other geometrical attributes are extracted from 4,096 patches (size 16×16) of four clusters. Features are then calculated by doubling the resolution up to five scales and then concatenating these features. Principal component analysis (PCA) [98] and LDA [99] are used to characterize images on the basis of morphological features, and a maximum of 84 percent accuracy is achieved. The second experiment focuses on 64×64 image block of each sample. Energy, contrast, and homogeneity are calculated from GLCM of the block by exploiting all possible options of distance $\mathrm{d} = 1, 2$ and angle $\theta = 0^0, 45^0, 90^0, 135^0$. Four-directional features are then concatenated, thereby resulting in an overall feature vector of size 24. LOO approach coupled with polynomial SVM is employed for classification. The reported classification accuracy is 90 percent.

Another example of hyperspectral analysis of colon biopsy samples is composite index (CI) measure-based SVM classification technique [56]. In this technique, spatial analysis on one selected spectral band of hyperspectral colon biopsy data is performed using circular local binary patterns (CLBP). CLBP(r, b) features are computed for 33 different combinations of radius $\mathrm{r} = 2, 3, \ldots, 12$ and number of neighbors $\mathrm{b} = 8, 12, 16$. For finding discriminative CLBP features, three measures are used based on separability and compactness of clustering. These measures are classification scatter index (C), rand index (R), and silhouette index (S). C is a measure of compactness of clusters formed by a set of features. It is calculated using following equation:

$$C = \sum_{j=1}^{m} p_j \sigma_j, \tag{9}$$

where p_j and σ_j, respectively, are the probabilities and standard deviations of the jth class.

S is a measure of similarity of each point to the points of its own cluster and to the points of other clusters. It is calculated as follows:

$$S = \sum_{i=1}^{N} \frac{\min[Db(i,k)] - Dw(i)}{\max[Dw(i), \min[Db(i,k)]]}, \tag{10}$$

where N is the total number of points. *Dw(i)* and *Dw(i,k)* are the average distances from the ith point to the other points in its own cluster, and to the points of another cluster k, respectively.

R measures how similar two clusters are to each other. It is calculated as follows:

$$R = \frac{\binom{N}{2}\sum_{i=1}^{g_1}\sum_{j=1}^{g_2} n_{ij}^2 - \frac{1}{2}\sum_{i=1}^{g_1}\left[\sum_{j=1}^{g_2} n_{ij}\right]^2 - \frac{1}{2}\sum_{j=1}^{g_2}\left[\sum_{i=1}^{g_1} n_{ij}\right]^2}{\binom{N}{2}}, \tag{11}$$

where N is the number of points in the image. G_1 and G_2 are the two partitions of data, having g_1 and g_2 clusters, respectively. n_{ij} is the number of points that belong to cluster i and cluster j of g_1 and g_2, respectively. CI is then calculated as weighted average of C, R, and S as follows:

$$CI = r_c C + r_r R + r_s S, \tag{12}$$

where r_c, r_r, and r_s are correlation coefficients for C, R, and S. CI depicted CLBP (5, 8) and CLBP (5, 12) as the two most promising features. Final classification involves LDA, PCA, and SVM. CLBP (5, 12) has maximum accuracy of 87.5 percent with SVM, whereas CLBP (5, 8) has maximum accuracy of 90.6 percent with PCA and SVM.

Maggioni et al. [57] presented a classification technique for discriminating normal, precancerous, and cancerous states of colon. In their work, tissue specimens are stained using H&E technique. Hyperspectral data is collected in the range of 440 to 700 nm, while setting the microscopic magnification factor to be 400×. Nuclei are extracted from the given samples, and then based on certain spectral features are assigned to one of the three classes. A total of 97.1 percent classification accuracy has been reported by the authors when nuclei have been extracted from all the samples. However, a gradual decrease in classification performance has been observed by the authors when nuclei have been extracted from smaller subsets of samples. Moreover, Chadded et al. [58] performed classification of multispectral colon biopsy images. In the first phase of this work, colon biopsy image is segmented by employing a modified version of classical Snake algorithm [100]. In the second phase, Haralick features of entropy, correlation, energy, homogeneity, and contrast are extracted from segmented portion of the image. Finally, images are classified into normal and malignant categories based on the extracted features. They achieved quite reasonable classification accuracy in discriminating different types of colon tissues.

Recently, Akbari et al. [59] proposed a method of colon cancer detection. They utilized a broad band light source to illuminate the tissue slide and a hyperspectral camera to capture wavelength bands from 450 to 950 nm. Twelve histo-pathological slides (three slides each for normal and malignant tissues of lung and lymph node) are used in their study. SVM is used to classify the given tissues. A total of 98.3 percent specificity and 96.2 percent sensitivity was observed for colon cancer data set.

3.4 Miscellaneous Techniques

There are a few other colon biopsy image classification techniques, which are not well established as other methods. Therefore, these techniques have been discussed in this section.

An example of such methods is visual analysis-based techniques, which simulate response of human receptive field operators, i.e., how human perceive the things in present scene and how neurons activate accordingly. One such study has exploited a few metrics, based on low-level process happening in human vision system [84]. Colon images of size 256×256 pixels are converted to 8-bit gray-scale images. Two features are calculated from these images, namely, total activation ratio (T) and orthogonal activation ratio (O). These parameters are supposed to mimic the neural activation happening in the brain by visualizing organizational structure of colon tissues. T is the ratio of maximum orientation response to the total neural activation. O is the ratio of maximum orientation response to the relevant orthogonal response. Maximum orientation response, used in the parameters, is measured over set of all orientations. Experiments proved the parameters to be good distinguishers of normal and malignant samples.

3.5 Gene Expression-Based Techniques

Gene expression profiling-based colon cancer detection is an active research area. There are usually three types of alterations a gene could undergo, i.e., over expression, suppression, and gene mutation. Such alterations have been exploited for detection of colon cancer, and significant research studies have been dedicated to this field. Genes are usually analyzed by using different variants of microarrays, like, Oligonucleotide and cDNA microarrays.

3.5.1 Oligonucleotide Microarrays

Oligonucleotide microarrays are created by synthesizing a particular Oligonucleotide in a solid surface based on an already defined spatial orientation. Oligonucleotide slides are scanned using nonfocal laser that analyzes different probes, and produces tiff images. Images are then analyzed to obtain level of gene expressions. Gene expressions are then used for classification.

To this end, Alon et al. employed a clustering algorithm on a data set of 6,500 gene expressions from 22 normal and 40 malignant colon tissues in their experiment. They found a set of 2,000 genes with the highest minimal intensity across samples [60]. These genes are supposed to be most discriminative compared to others in the data set. In 2007, Grade et al. [61] worked on gene data of 73 malignant and 30 normal patients, and found 17 discriminative genes among the data. Moreover, Yajima et al. [62] analyzed gene expression profiles of 43 patients (23 curable cancer patients—stage A-C, 14 early cancer patients—stage A-B, 5 right-sided cancer patients—stage D) in their research study. Gene expression profiles are obtained from the feces and peripheral blood. Three (PAP, REG1A, and DPEP1) and six (SEPP1, RPL27A, ATP1B1, EEF1A1, SFN, and RPS11) most distinctive genes are identified, respectively, in the samples of peripheral blood and feces. This set of nine genes has proven to be able to accurately identify 78 percent of stage A-C, 71 percent of stage A-B, and 80 percent of stage D patients. Likewise, Kim et al. [63] worked on a data set of five serrated adenomas and five normal colon mucosa samples, and identified 124 discerning genes capable of distinguishing the samples in an effective manner.

Later, Venkatesh et al. [64] proposed another method of colon cancer detection. Kent Ridge colon cancer data set has been used that contains 2,000 gene expressions with highest minimal intensity across 62 tissues. In this work, dimensionality of data is reduced by using chi-squared measure, and 135 out of 2,000 genes are selected after ranking. A recurrent neural network with context layer, called FEJ neural network, is used for classification. The reported classification accuracy is 94.44 percent for this work that is better than that of other classifiers such as naive Bayes, classification and regression tree (CART) and random tree by 10.23 percent approximately. Kulkarni et al. [65] proposed an evolutionary algorithms-based method for automatic detection of colon cancer. In this work, t-statistic and mutual information are employed for selection of discriminative genes among a given pool of genes. Genetic programming and decision trees are employed as classifiers, and data is divided into normal and malignant samples based on top 10 and top 20 selected genes. Result revealed that mutual information-based feature selection together with genetic programming is the most effective solution compared to other combinations.

Recently, Lee et al. [66] proposed a colon cancer detection technique. In this study, newly proposed neural network-based finite impulse response extreme learning machine (FIR-ELM) [101] is employed. The FIR-ELM algorithm performs classification based on single hidden layer feed forward neural network (SLFN). In SLFN, well-known filtering methods, like, finite length low-pass filtering, high-pass filtering, and band-pass filtering are employed to train the input weights in the hidden layer of SLFN to extract features from the data set. These features are then used to classify the given colon samples. Further, Tong et al. proposed an ensemble of SVM classifiers-based method of colon cancer detection [67]. In this work, 50 gene expressions are selected using top scoring pair method, and linear SVM classifiers are trained on those pairs. GA is employed to select such an optimal combination of SVM base classifiers, which yields maximum possible performance. They investigated the effectiveness of their technique on several binary class and multiclass gene expression data sets including one on colon cancer. They reported classification accuracy of 90.30 percent with colon data set.

3.5.2 cDNA Microarrays

Like Oligonucleotide, considerable research has been carried out by using cDNA microarrays. In 1999, Backert et al. [68] have utilized 588 genes, obtained from three classes (normal, mucinous and nonmucinous) of colon tissues in their research work. Mucinous and nonmucinous are two phenotypes of colon carcinoma that bear morphological as well as genetic differences. Cell lines are prepared for each colon category. RNA is extracted from cultured cell lines using RNAZol which after being polyadenylated is enriched using magnetic dynabeads, while maintaining quality with agarose gel electrophoresis. Duplicate copies of the genes are spotted on a nylon membrane. Membrane is then hybridized with labeled cDNA probes, prepared by reverse transcription from $1\mu g$ $\mathrm{polyA} + \mathrm{mRNA}$. Ten alterations are detected in cell lines of malignant colon compared to normal one. The reported classification accuracy is more than 50 percent.

Bianchini et al. [70] worked on gene expressions of 25 malignant and 13 normal samples, and identified

584 discriminating genes. Li et al. [71] have used GA to identify discriminative genes, and achieved classification accuracy of 94.1 percent by using KNN classifier. Similarly, Chen and Li [72] used multiple kernel SVM (MK-SVM) technique where multiple kernels are described as the convex combination of the single feature basic kernels. Algorithm was tested on two gene expression data sets: leukemia data and colon data, and more than 90 percent classification success was achieved for both the data sets. Strive continued in this field, and Shon et al. [73] proposed working in the frequency domain. They used wavelet transform to reduce feature space, and obtained 92 percent accuracy on colon data with probabilistic neural network (PNN).

3.6 Laser Induced Fluorescence (Blood Serum Analysis)-Based Techniques

Cancer changes chemical composition of different ingredients in blood serum. Consequently, Raman spectrum of malignant serum heavily deviates from its normal counterpart. Techniques, based on laser-induced fluorescence and Raman spectroscopy, exploit such differences in blood serum and resultant Raman spectra for detection of colon cancer.

In one of such studies, veins of patients are separated in a segregator at a speed of 3,000 rot/min for 10 min [74]. Then upper serum is sucked, and developed samples are hermitically refrigerated. Double monochromator equipped with PMT is used to collect spectra. Samples are directly exposed to Ar-ion laser source operating at 488 nm or 514.5 nm and resultant Raman radiation is put into monochromator. Spectrum is then amplified by using a lock-in amplifier, and saved in computer for further use. The reported classification accuracy is 83.5 percent for a data set of 65 samples. Analysis revealed that normal serum showed three well-distinguished peaks in the spectrum, whereas cancerous serum showed no sharp peak. Fluorescence spectroscopy also forms basis of an 88 percent accurate probability-based algorithm for detection of colon cancer [75]. Specimen collection is quite lengthy; patients are ingested with colyte solution before colonoscopy. Distal tip of spectro fluorimeter, surrounded by bundle of nine collection fibers, is placed multiple times in light contact with polyp and three spectra are collected for each placement. After this, a fluorescence spectrum is collected from a normal area of colon, approximately 1 cm far from the polyp. Postprocessing is applied on the fluorescence spectra to refine it. Fluorescence data is divided into modeling and validation data sets. Modeling data is used to develop an algorithm that could discriminate colon tissues. Spectral regions, containing the most useful diagnostic information, are identified. Diagnostic parameters for these regions are identified, and their probability distribution is used to construct a diagnostic algorithm. The algorithm is first applied to the modeling data, and then is blindly tested on validation data to determine its accuracy. The reported classification accuracy for normal, hyperplastic, and adenomatous samples is 97, 50, and 84 percent, respectively.

4 Comparison

This section provides a detailed comparison of colon cancer detection categories, and of multiple techniques within each category. This section also summarizes equipment and the data set used for validating these techniques. Pros and cons of various techniques are also a part of this section.

4.1 Comparison of Equipment and Data Set

Equipment and data set play a pivotal role in determining overall trustworthiness of any technique. Advanced data acquisition equipment and larger data sets usually ensure the reliability of a technique. Table 2 summarizes the equipment used in data acquisition, equipment settings used for capturing images/spectra, and data set for the techniques explained in Section 3.

4.2 Performance Comparison

It is difficult to compare different techniques as each one of them uses its own data set and equipment. However, we have compared techniques within each category based on a few parameters such as accuracy, data set and its acquisition method, cancer detection and grading capability, and parameter tuning. A brief comparison of different cancer detection categories is in Table 3. Table 4 provides detailed comparison of colon cancer detection techniques within each category. Pros and cons of individual techniques are highlighted in Table 5.

Accuracy is the most promising parameter to measure effectiveness of a technique. Table 3 demonstrates low variability in terms of accuracy. However, OO texture analysis-based techniques [44], [45], [46], [47], [48], [49], [50], [51], [52] go ahead of others. Primary reason of better accuracy is incorporation of background knowledge about tissues organization into the segmentation/classification process. Contrary, laser-induced fluorescence-based techniques [74], [75], [76], [77], [78], [79], [80], [81], [82], [83] have smaller accuracy, because equipment is quite delicate and a minute human error leads to wrong results. Texture [35], [36], [37], [38], [39], [40], [41], spectral [53], [54], [55], [56], [57], [58], [59], and gene analysis-based techniques [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73] show moderate accuracy. Second factor is availability of equipment. Nikon Coolscope microscope [102] and CCD cameras have been used in visual analysis and texture analysis-based techniques. This equipment is easy to use and is easily accessible to histopathologists. Fluorescence, gene analysis, and hyperspectral analysis-based techniques are hard to practice because equipment is quite delicate and is not easily accessible to histopathologists. Cancer detection and grading is another factor. All the techniques have capability to detect cancer except OO texture analysis-based segmentation techniques [44], [45], [46], [47], [48], [49]. There is no cancer grading capability in different techniques except the detection of two phenotypes in Backert et al.'s work [68], hyperplastic and adenomatous stages in Cothern et al.'s work [75], and four Duke's stages in Yajima's work [62]. Parameter tuning is a major challenge in OO texture, fluorescence, simple texture, and spectral analysis-based techniques. Most of the techniques within these categories need manual/partial adjustment of parameters. Contrary, gene and visual analysis-based techniques do not need parameter tuning.

4.2.1 Comparison: OO Texture-Based Techniques

Table 4 demonstrates that OO texture analysis suits to the segmentation of colon biopsy images as depicted by the

TABLE 2
Comparison of Equipment, Data Set, and Image Acquisition

Technique	Equipment	Dataset	Mag. / Wavelength (nm)	Staining
Textural Analysis Based Techniques				
Esgiar et al. [35]	Light Microscope, CCD camera	44 normal, 58 malignant images	40x	IH
Esgiar et al. [36]	Light Microscope, CCD camera	44 normal, 58 malignant images	40x	IH
Esgair et al. [37]	Light Microscope, CCD camera	44 normal, 58 malignant images	40x	IH
Esgiar et al. [38]	Light Microscope, JVC TK-1280E camera	44 normal, 58 malignant images	40x	IH
Kalkan et al. [39]	-----	8000 images(patches), (36 patients)	-----	H&E
Jiao et al. [40]	-----	30 normal, 30 malignant images	----	----
Object-oriented Texture Analysis Based Techniques				
Tosun et al. [44]	Nikon Cool scope Microscope	16 images	5x	H&E
Tosun et al. [45]	Nikon Cool scope Microscope	16 images	5x	H&E
Demir et al. [46]	Nikon Cool scope Microscope	72 images (36 patients)	20x	H&E
Tosun et al. [47]	Nikon Cool scope Microscope	150 images	5x	H&E
Simsek et al. [48]	Nikon Cool scope Microscope	200 images	5x	H&E
Rathore et al. [49]	Nikon Cool scope Microscope	100 images (68 patients)	4x, 5x, 10x, 40x	H&E
Altunbay et al. [50]	Nikon Cool scope Microscope	213 images (58 patients)	20x	H&E
Ozdemir et al. [51]	Nikon Cool scope Microscope	3236 images (258 patients)	20x	H&E
Ozdemir et al. [52]	Nikon Cool scope Microscope	3236 images (258 patients)	20x	H&E
Hyperspectral Analysis Based Techniques				
Rajpoot et al. [54]	Nikon Biophot Microscope, CCD camera	32 samples	40x / 450-640	H&E
Masood et al. [55]	CRI Nuance Microscope, CCD camera	2 slides [1]	400x	H&E
Masood et al [56]	Nikon Biophot Microscope, CCD camera	32 samples	40x /440-700	H&E
Maggioni et al. [57]	Nikon Biophot Microscope, CCD camera	59 samples	440-700	H&E
Chadded et al. [58]	-----	45 images	500-650	-----
Akbari et al. [59]	Macroscopic optical histopathology	12 slides	450-950	-----
Gene Analysis Based Techniques				
Alon et al. [60]	Oligonucleotide microarrays	65000 genes(22 normal, 40 malignant)	-----	-----
Grade et al. [61]	Oligonucleotide microarrays	17 genes (30 normal, 73 malignant)	-----	-----
Yajima et al. [62]	Oligonucleotide microarrays	Stages: A-C (23), A-B (14), D (5)	-----	-----
Kim et al. [63]	Oligonucleotide microarrays	124 genes (5 normal, 5 malignant)	-----	-----
Venktesh et al. [64]	Oligonucleotide microarrays	2000 genes (22 normal, 40 malignant)	-----	-----
Kulkarni et al. [65]	Oligonucleotide microarrays	2000 genes (22 normal, 40 malignant)	-----	-----
Lee et al. [66]	Oligonucleotide microarrays	2000 genes (22 normal, 40 malignant)	-----	-----
Tong et al. [67]	Oligonucleotide microarrays	2000 genes (22 normal, 40 malignant)	-----	-----
Backert et al. [68]	cDNA microarrays	588 genes	-----	-----
Bianchini et al. [70]	cDNA microarrays	584 genes (25 malignant, 13 normal)	-----	-----
Laser-Induced Fluorescence Based Techniques				
Li et al. [74]	Double Monochromator equipped with PMT	65 patients	520-640,540- 560, 500-620 /510- 530	-----
Cothren et al. [75]	Spectro Fluorimeter with 9 collection fibers	172 samples	400- 700	-----

1= each slide has several microdots, ----- = no information available or the entry is irrelevant

TABLE 3
Comparison of Colon Cancer Detection Categories

	Object Oriented Texture analysis	Laser-Induced	Texture Analysis	Spectral Analysis	Genes analysis	Visual Analysis
Accuracy	**99(segmentation), 92.21(classification)**	**88**	**96.67**	**97.10**	**98.33**	**----**
Equipment (Easily Available)	√	×	√	×	×	√
Ease of Use for Histopathologists	√	×	√	×	×	√
Cancer Detection	--	√	√	√	√	√
Cancer Grading	×	--	×	×	--	×
Automatic Parameter Tuning	--	--	--	--	√	√

√=fully satisfy, --=partially satisfy, x=do not satisfy criteria, Accuracy of the best performing technique within each category is given.

superior segmentation results compared to classification. Second, within the segmentation results, graph-based segmentation [46], [47], [48] seems better compared to object-based segmentation [44], [45]. Possible reason of better accuracy is that graphs and resultant features represent tissue components more realistically compared to simple object/pixel-based features. Except initial two techniques [44], [45] in this category, all other techniques

TABLE 4
Comparison of Colon Cancer Detection Techniques

	Accuracy	Data Acquisition/ Dataset	Cancer Detection	Cancer Grading	Parameter Tuning
Textural Analysis Based Techniques					
Esgiar et al. [35]	90.20%	Smaller dataset	Yes	No	Manual tuning
Esgiar et al. [36]	85.00%	Smaller dataset	Yes	No	Manual tuning
Esgiar et al. [37]	94.10%	Smaller dataset	Yes	No	Manual tuning
Esgiar et al. [38]	94.10%	Smaller dataset	Yes	No	Manual tuning
Kalkan et al. [39]	75.15%	Larger dataset	Yes	No	Partially automatic
Jiao et al. [40]	96.67%	Smaller dataset	Yes	No	Fully automatic
Object-Oriented Texture Analysis Based Techniques					
Tosun et al. [44]	94.80%	Smaller dataset	No	No (segmentation)	Manual tuning
Tosun et al. [45]	86.50%	Smaller dataset	No	No (segmentation)	Partially automatic
Demir et al. [46]	90.62%	Smaller dataset	No	No (segmentation)	Partially automatic
Tosun et al. [47]	99.00 %	Larger dataset	No	No (segmentation)	Partially automatic
Simsek et al. [48]	94.90%	Larger dataset	No	No (segmentation)	Manual tuning
Rathore et al. [49]	76.99(4x),85.77(5x), 88.98(10x),76.54(40x)	Larger dataset	No	No (segmentation)	Fully automatic (GA)
Altunbay et al. [50]	82.65%	Larger dataset	Yes	No	Partially automatic
Ozdemir et al. [51]	90.66% (average)	Much larger dataset	Yes	No	Partially automatic
Ozdemir et al. [52]	92.21%	Much larger dataset	Yes	No	Partially automatic
Hyperspectral Analysis Based Techniques					
Rajpoot et al. [54]	87.5%	High resolution data cube, Visible range: 450-640nm	Yes	No	Manual Tuning
Masood et al. [55]	90%	Short process, common staining method	Yes	No	Fully Automatic
Masood et al [56]	90.6%	Low resolution data cube, Visible range: 440-700nm	Yes	No	Fully Automatic
Maggioni et al. [57]	97.1%	440-700nm	Yes	No	Fully Automatic
Chadded et al. [58]	-----	500-650nm	Yes	No	Partially automatic
Akbari et al. [59]	97%	450-950nm	Yes	No	-----
Gene Analysis Based Techniques					
Venktesh et al.[64]	94.4%	Kent-Ridge data set, larger data set	Yes	No	Fully Automatic
Backert et al. [68]	>50%	Data itself prepared, lengthy process, data set is small	Yes	Detects 2 phenotypes	Fully Automatic
Kulkarni et al. [65]	98.33	Kent-Ridge data set, larger data set	Yes	No	Partially automatic
Lee et al. [66]	76.85	Kent-Ridge data set, larger data set	Yes	No	Partially automatic
Tong et al. [67]	90.03%	Kent-Ridge data set, larger data set	Yes	No	Partially automatic
Laser-Induced Fluorescence Spectroscopy Based Techniques					
Li et al. [74]	83.50%	Small data set	Yes	No	Fully Automatic
Cothren et al. [75]	88%	Large dataset	Yes	Detects hyperplastic and adenoma	Manual tuning

----- = no information available or the entry is irrelevant

have been tested on larger data sets. Data acquisition is same because samples in all the techniques have been stained using H&E technique, and Nikon Coolscope has been used for capturing images. Segmentation techniques [44], [45], [46], [47], [48], [49] only segment images, whereas classification techniques classify samples into different classes. There is no cancer grading capability in either of the techniques. Parameter tuning is manual in a few techniques [44], [48]. For instance, several parameter such as circle radius, window size, and merge threshold are manually adjusted in OOSEG [44]. Conversely, parameter tuning is partially automatic in others [45], [46], [47], [50], [51], [52]. In these techniques, authors have used separate data set for finding optimal values of parameters. There is only one technique [49] that provides automatic adjustment of parameters through GA.

4.2.2 *Comparison: Spectral Analysis-Based Techniques*

Hyperspectral analysis-based techniques share some common advantages and disadvantages, and are poles apart in various respects. Such similarities and differences have been presented in Table 4. All the techniques yield reasonable classification results. However, Maggioni's work [57] leads others in terms of accuracy. This is primarily due to larger visible range, established to acquire hyperspectral data. Like many others, all the hyperspectral analysis-based techniques only distinguish normal and malignant samples, and are silent about cancer grading.

TABLE 5
Pros and Cons of Colon Cancer Detection Techniques

Technique	Pros	Cons
Textural Analysis Based Techniques		
Esgiar et al. [35]	-Simple and straightforward. -Quite common statistical parameters are exploited. -Easy to use for histopathologists.	-Quite delicate and lengthy data acquisition method.
Esgiar et al. [38]	-Simple and straightforward parameters are used. -Quite less computational complexity. -Easy to use for histopathologists.	-Little improvement because fractal dimensions are highly correlated with entropy and correlation.
Jiao et al. [40]	-Simple to use for histopathologists. -Simple and straightforward approach; -Image data is easily available in hospitals. -Small and compact feature set is used.	-Smaller dataset.
Object oriented Texture Analysis Based Techniques		
Tosun et al. [44]	-Simple to use for histopathologists. -Image data is easily available in hospitals.	-Can only segment heterogeneous images of 5x microscopic magnification. -Change in magnification degrades the results. -Parameters are manually adjusted. -Smaller data set.
Tosun et al. [45]	-Simple to use for histopathologists. -Image data is easily available in hospitals. -Manual adjustment of parameters has been eliminated to a huge extent.	-Can only segment heterogeneous images of 5x microscopic magnification. - Change in magnification degrades the results. -Smaller dataset.
Demir et al. [46]	-Simple to use for the histopathologists. -Image data is easily available. -Introduced graphs for the segmentation of colon biopsies for the first time.	-Can only segment images of 20x microscopic magnification. -Partial automation of system parameters. -Computationally expensive.
Tosun et al. [47]	-Simple to use for histopathologists. -Much accurate segmentation (99%) with no restriction on number of segmented regions.	-Can only segment heterogeneous images of 5x microscopic magnification. - Change in magnification degrades the results. -Computationally expensive.
Rathore et al. [49]	-Simple to use for the histopathologists. -Applicable to four different magnification factors. -Parameter tuning is totally automatic through GA. -Independent of orientation of slide on microscope while capturing images.	-Computationally expensive as computation of parameters through GA, and ellipse finding in four angular directions take considerable time.
Hyperspectral Analysis Based Techniques		
Rajpoot et al. [54]	-Caters local as well as global level changes. -Simple and straightforward.	-Independent components are used instead of full 3D data which limit the accuracy. -Spectral data is not easily available to histopathologists.
Masood et al. [55]	-Simple and straightforward approach. -Dimensionality reduced compact spectral data is used. -Compact feature vector is used.	-Spectral data is not easily available to histopathologists.
Masood et al [56]	-CLBP features, along with three clustering measures, are used which serve as good discriminators of samples. -Much smaller and compact feature vector. -Simple and straightforward.	-Only one band is used compared to full 3D data which limits accuracy of the scheme. -Spectral data is not easily available to histopathologists.
Maggioni et al. [57]	-Simple and straightforward technique. -Spectral features are used which serve as good discriminators of colon samples. -Computationally tractable.	-Spectral data is not easily available to histopathologists.
Chadded et al. [58]	-Introduced valuable modification to the Snake algorithm. -Only three features are used to classify the segmented regions, which are much lesser than many other schemes.	-Snake needs to be initially demarcated. -Dataset is quite small. -Quantitative result and comparisons are not provided.
Akbari et al. [59]	-Simple and straightforward approach. -Performs equally well for lung and colon cancer. -No complex parameters are involved.	-Dataset is quite small.
Gene Analysis Based Techniques		
Venktesh et al. [64]	-Reduced feature set has been used. -Comprehensive gene expression database has been used.	-Synthesis of Oligonucleotide is lengthy process. -Difficult to practice for histopathologists.
Backert et al. [68]	-Grading capability of 2 phenotypes of carcinoma.	-Difficult to practice for histopathologists. -Lengthy and delicate data acquisition process.
Kulkarni et al. [65]	-Reduced feature set has been used. -Comprehensive gene expression database has been used. -Benefits from capabilities of evolutionary algorithms.	- Difficult to practice for histopathologists. -Computationally expensive.
Lee et al. [66]	-Comprehensive gene expression database has been used. -Time series analysis serves well for colon dataset. -Technique performs equally well for given two datasets.	- Difficult to practice for histopathologists. - Weight updation is computationally expensive. - Selection of number of neurons for hidden layer is a major challenge.
Tong et al. [67]	-Comprehensive gene expression database has been used. -Performance has been enhanced due to ensemble.	- Difficult to practice for histopathologists. - Computationally expensive.
Laser-Induced Fluorescence Spectroscopy Based Techniques		
Li et al. [74]	-Easy to detect peaks and classify images for pathologist. -Amplified spectrum is used. -Several wavelength ranges are observed.	-Quite complicated data acquisition. -Equipment is not easily available to histopathologists.
Cothren et al. [75]	-Extensive data corrections steps are applied before further processing. -Average spectra are used instead of single spectra. -Detects two stages of hyper plastic and adenomatous.	-Quite complicated data acquisition. -Equipment is not easily available to histopathologists.
Visual Analysis Based Techniques		
Todman et al. [85]	-Simple to use. -Quite less computational complexity.	-Not reliable. -No quantitative results have been provided.

4.2.3 *Comparison: Gene Expression-Based Techniques*

Ultimate aim of gene expression analysis-based techniques is to classify gene expression profiles into normal and malignant categories. These techniques usually select multiple gene expressions among a given pool to classify given colon samples. Researchers have tried to find out such common genes which are significantly expressed in a large number and diverse colon cancer samples. In this context, Sanz-Pamplona et al. [103] have investigated the discerning capability of 31 different gene expressions in 11-genes-based colon data sets. But, they could find only one gene that differentiates samples in 11 given data sets

with more than 65 percent classification accuracy. Therefore, they concluded that there is no single gene which can distinguish malignant samples in diverse data sets. Rather, a combination of genes should be used. Among the gene analysis-based techniques presented in Section 3.5, Oligoneuclotide-based techniques are rich in terms of data set. Larger data set is used in these techniques. On the other hand, cDNA-based techniques have been tested on smaller data sets. All the techniques have cancer detection capability. There are only two techniques which assign cancer grades; Backert's work [68] that divides samples into two phenotypes, and Yajima's work [62], which assigns quantitative cancer grades according to Duke's scale. Further, there is no manual adjustment of parameters in either of the techniques.

4.2.4 Comparison: Texture Analysis-Based Techniques

These techniques have exploited texture features for detection of colon cancer. The work of Jiao et al. [40] enjoys a slight performance advantage over others. Possible reason of higher accuracy is the use of compact and information rich hybrid feature vector. All the techniques [35], [36], [37], [38], [39], [40], [41] classify colon samples into normal and malignant classes but none of them distinguishes cancer grades. Data set is smaller in a few studies [35], [36], [37], [38], [40], and vice versa [39], [41]. Kalkan et al.'s work [39] is partially automatic, and Jiao et al.'s technique [40] is fully automatic. On the other hand, remaining techniques need manual adjustment of system parameters such as subimage size and area threshold needs to be manually adjusted, respectively, in [38] and [35].

4.2.5 Comparison: Laser-Induced Fluorescence-Based Techniques

These techniques analyze the Raman spectrum of blood serum. Three well-distinguished peaks are indicator of normal blood serum, whereas irregular peaks or absence of peak shows cancer. There is small difference of accuracy in both the techniques. Primary reasons of better accuracy in spectro fluorimeter-based technique [75] are the postprocessing steps. Data set in spectro fluorimeter-based technique [75] is large (172 samples), but data acquisition process is pretty lengthy and delicate. Placement and removal of distal tip in light contact with the polyp needs extreme care. Contrary, data set is smaller (65 patients) in [74] but data acquisition is simple and straightforward. Cancer detection process is totally automatic in monochromator-based technique [74]. However, optimal values of system parameters are calculated by analyzing their probability distribution in [75]. Both the techniques detect normal and malignant samples. However, spectro fluorimeter-based technique distinguishes between two precancerous stages as well.

5 Experimental Evaluation

In this section, most of the techniques presented in Section 3 have been evaluated. The techniques within different categories work in different domains, therefore, it is not possible to test all of them by using similar data set. However, techniques within each category have been evaluated on the same data set. In this connection, three data sets have been prepared/acquired, and the different techniques have been implemented in Matlab.

5.1 Data Set

Data Set-I has been used to assess gene expression-based techniques. Data Set-II has been used to test OO texture analysis-based segmentation techniques, and data set-III has been used to test texture, visual, and OO texture analysis-based classification techniques.

Data Set-I. This data comprises two standard colon cancer data sets (Kent Ridge data set [104] and BioGPS data set [105]). Kent Ridge data set contains 62 samples (22 normal and 40 malignant). Dimensionality of Kent Ridge data set is 2,000. BioGPS data set comprises 131 samples (37 normal and 94 malignant). Dimensionality of BioGPS data set is 3.

Data Set-II. This data set comprises 100 biopsy images taken from 68 randomly selected patients from the pathology department of Rawalpindi Medical College. Biopsy samples, comprising 5-6-μm thick tissue section, have been stained with H&E. Nikon Biophot microscope has been used with four different magnification factors $(4\times, 5\times, 10\times, 40\times)$ for capturing images. Each image has a spatial resolution of 800×600. Images comprise normal, malignant or both tissue types.

Data Set. Data Set-III has been prepared by capturing 174 RGB images under lens magnification factor of $10\times$ from the histopathological slides of 68 patients (as used in data set-II). Out of 174 images, 92 are malignant and 82 are benign. The images have varying size starting from minimum of 67×160 to a maximum of 704×376.

5.2 Segmentation Experiment

The segmentation experiments have been performed on the data set-I and visual results have been obtained for various colon biopsy image segmentation techniques, as presented in Fig. 6.

Figs. $6a_1$, $6a_2$, $6a_3$, $6b_1$, $6b_2$, and $6b_3$, respectively, represent normal and malignant colon samples captured at $40\times$ magnification factor. It is clearly evident that multilevel segmentation-based technique [48] clearly outclasses the remaining two. This is largely due to its ability to ensemble diverse segmentation results achieved by segmenting diverse graphs. Furthermore, we have quantitatively evaluated the segmentation results using well-known performance measures such as accuracy, sensitivity, specificity, MCC, and F-Score.

Accuracy is a measure of overall usefulness of the classification technique. It can be calculated using following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100, \qquad (13)$$

where true positive (TP), true negative (TN) are the number of correctly classified positive and negative samples. False positive (FP) and false negative (FN) are incorrectly classified samples.

Sensitivity and specificity, respectively, are used to calculate ability of a classifier to recognize patterns of

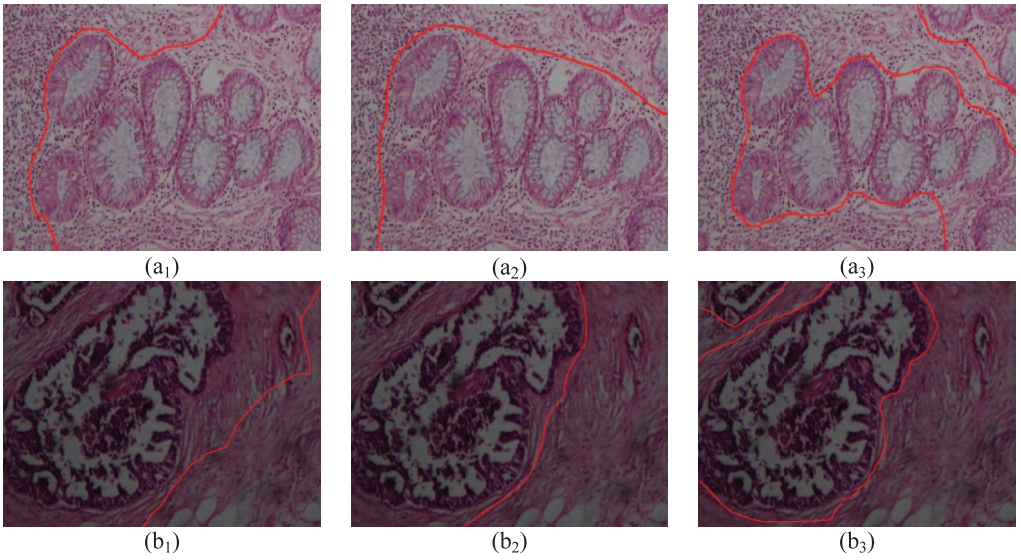


Fig. 6. Segmentation results for two images using (a_1, b_1) OOSEG [44], (a_2, b_2) GRLM [47], and (a_3, b_3) multilevel segmentation [48].

positive and negative classes. They can be obtained using the following expressions:

$$Sensitivity = \frac{TP}{TP + FN}, \tag{14}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{15}$$

MCC serves as a measure of classification in binary class problems. Its value ranges from -1 to $+1$. $+1$ means classifier is always right, whereas -1 means classifier always commits a mistake. 0 means random prediction. MCC can be calculated using the following expression:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FN)(TP + FP)(TN + FN)(TN + FP))}}. \tag{16}$$

F-score is a weighted average of precision and recall values. It can be calculated by using expression (19). Its value ranges between 0 and 1, where 0 is the worst score and 1 is the best.

$$Precision = \frac{TP}{TP + FP}, \tag{17}$$

$$Recall = \frac{TP}{TP + FN}, \tag{18}$$

$$Fscore = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{19}$$

Quantitative results for various segmentation techniques have been presented in Table 6.

Results in Table 6 reinforce the conclusions drawn from Fig. 6, wherein we see that multilevel partitioning-based technique [48] outclasses other techniques [44], [45], [47] in terms of most of the performance evaluation parameters.

5.3 Classification Experiment

Likewise, the classification experiments have been conducted on the aforementioned data sets (Data Set I and Data Set III) and performance of most of the classification techniques has been evaluated. Visual classification result of five images (see Fig. 7) are given in Table 7.

Table 7 reveals that image in Fig. 7d is misclassified by most of the techniques. This might be due to the fact that image presents precancerous stage, in which tissues are slightly deformed. So, it is hard to detect cancer. Quantitative results for classification have been given in Table 8.

6 Conclusion

Traditionally, colon cancer is diagnosed using microscopic tissue analysis. However, the process is subjective, and may lead to interobserver variation in grading. Further, factors such as tiredness, experience, and workload of pathologist also affect the diagnosis. These vulnerabilities in the manual

TABLE 6
Quantitative Results for Object-Oriented Texture Analysis-Based Segmentation Techniques

Colon biopsy image segmentation Technique	Ref.	Performance measures				
		Accuracy	Sensitivity	Specificity	MCC	F-Score
Tosun et al.	[44]	81.29±1.23	0.94±0.01	0.73±0.02	0.80±0.03	0.79±0.01
Tosun et al.	[45]	82.45±1.11	0.96±0.02	0.78±0.03	0.85±0.01	0.83±0.02
Tosun et al.	[47]	83.50±1.02	0.93±0.02	0.80±0.03	0.82±0.02	0.81±0.01
Simsek et al.	[48]	85.77±0.97	0.88±0.01	0.85±0.01	0.86±0.02	0.83±0.02

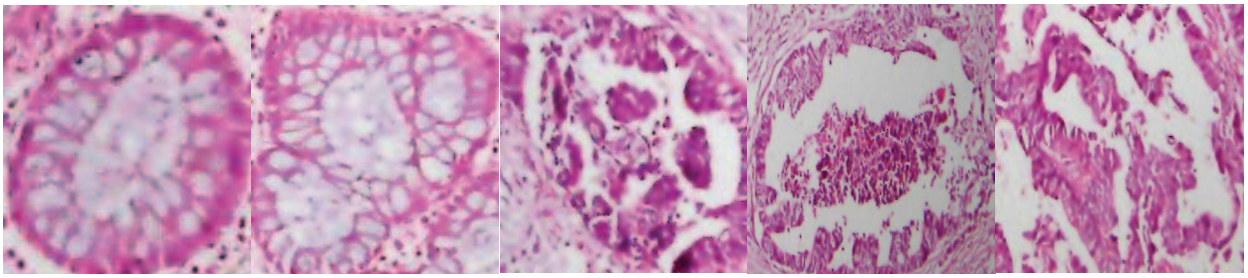


Fig. 7: (a) and (b) Normal and (c)-(e) malignant colon biopsy images.

TABLE 7
Classification Results for Images Given in Fig. 7 by Using Various Colon Classification Techniques (N = Normal, M = Malignant)

Classification technique	Ref.	Labels				
		Fig 7 (a)	Fig 7(b)	Fig 7(c)	Fig 7(d)	Fig 7(e)
Esgiar et al. (KNN)	[35]	N	N	N	M	N
Esgiar et al. (LDA)	[35]	N	M	M	N	M
Esgiar et al.	[36]	M	N	M	N	M
Esgiar et al.	[37]	N	N	M	M	M
Esgiar et al.	[38]	N	N	M	M	M
Kalkan et al.	[39]	N	N	M	M	M
Jiao et al.	[40]	N	N	M	N	M
Todman et al.	[85]	N	N	M	N	N

TABLE 8
Quantitative Results for Various Colon Classification Techniques (KR = Kent Ridge Data Set, BG = BioGPS Data Set)

Colon classification technique		Ref.	Performance measures				
			Accuracy	Sensitivity	Specificity	MCC	F-Score
Texture analysis based	Esgiar et al. (KNN)	[35]	77.75±1.02	0.88±0.01	0.71±0.01	0.58±0.02	0.76±0.01
	Esgiar et al. (LDA)	[35]	86.00±1.32	0.89±0.09	0.95±0.04	0.54±0.07	0.84±0.12
	Esgiar et al.	[37]	73.18±1.78	0.82±0.02	0.67±0.02	0.48±0.03	0.71±0.02
	Esgiar et al.	[38]	73.18±1.78	0.82±0.02	0.67±0.02	0.48±0.03	0.71±0.02
	Kalkan et al.	[39]	83.22±1.01	0.80±0.01	0.81±0.04	0.85±0.01	0.79±0.03
	Jiao et al.	[40]	81.15±0.65	0.83±0.01	0.80±0.02	0.81±0.02	0.82±0.03
Gene expression Analysis based	Venkatesh et al. (KR)	[64]	94.10±1.54	0.93±0.02	0.94±0.03	0.91±0.01	0.90±0.02
	Venkatesh et al. (BG)	[64]	96.62±1.23	0.97±0.01	0.97±0.02	0.98±0.01	0.97±0.01
	Kulkarni et al. (KR)	[65]	98.33±1.00	0.99±0.01	0.98±0.01	0.97±0.02	0.97±0.02
	Kulkarni et al. (BG)	[65]	98.45±0.95	0.99±0.01	0.99±0.01	0.98±0.01	0.98±0.02
	Lee et al. (KR)	[66]	76.85±1.56	0.77±0.05	0.79±0.04	0.70±0.03	0.74±0.03
	Lee et al. (BG)	[66]	80.23±2.13	0.79±0.04	0.81±0.05	0.77±0.02	0.81±0.01
	Tong et al. (KR)	[67]	90.32±1.11	0.82±0.02	0.95±0.02	0.79±0.03	0.86±0.02
	Tong et al. (BG)	[67]	93.55±1.28	0.86±0.01	0.98±0.01	0.86±0.02	0.90±0.02
	Li et al. (KR)	[71]	89.01±1.02	0.88±0.02	0.90±0.01	0.87±0.03	0.88±0.02
	Li et al. (BG)	[71]	91.25±1.03	0.89±0.01	0.93±0.02	0.90±0.01	0.88±0.02
Miscellaneous	Todman et al.	[85]	88.32±1.59	0.88±0.02	0.86±0.03	0.90±0.01	0.83±0.05

process result in need of an automatic colon cancer detection mechanism. In this context, several colon cancer detection techniques have been proposed. In this survey, we have divided these techniques into five major categories: OO texture, spectral, spatial, basic texture, serum, and gene analysis-based techniques. A larger subset of these techniques has been summarized in this paper. Additionally, an extensive comparison of various colon cancer detection categories and of multiple techniques within each category has also been provided. Most of the techniques have been implemented in Matlab, and tested on unified data set. Analysis reveals that simple texture and OO texture analysis-based techniques are better compared to other approaches owing to their ease of use for histopathologits, easy access to the equipment, and superior results. Though our preliminary survey is quite promising, there is still a lot more to be done. First of all, a few other performance measures may be introduced in the comparison. Second, there should be a separate study focusing on parameter tuning of these techniques.

ACKNOWLEDGMENTS

This work was supported by the Higher Education Commission of Pakistan under the Indigenous PhD Scholarship Program as per award no. 117-7931-Eg7-037. The authors would like to thank Mr. Imtiaz Ahmad Qureshi (Assistant Professor, Histopathology Department, Rawalpindi Medical College) for providing data and a relevant expert opinion. They are also grateful to Miss Terese Winslow (http://www.teresewinslow.com/) for providing Fig. 2.

REFERENCES

[1] Ashiya, "Notes on the Structure and Functions of Large Intestine of Human Body," http://www.preservearticles.com/201105216897/notes-on-the-structure-and-functions-of-large-intestine-of-human-body.html, Feb. 2013.

[2] A. Mandal, "What Does the Large Intestine Do?" http://www.news-medical.net/health/What-Does-the-Large-Intestine-Do.aspx, Feb. 2013.

[3] R.S. Houlston, "Molecular Pathology of Colorectal Cancer," *J. Clinical Pathology,* vol. 54, pp. 206-214, 2001.

[4] Colon Cancer Alliance, "Colon Cancer Risk Factors," http://www.ccalliance.org/colorectal_cancer/riskfactors.html, Dec. 2012.

[5] PubMed Health, "Colon Cancer," http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001308/, Feb. 2013.

[6] G.D. Thomas et al., "Observer Variation in the Histological Grading of Rectal Carcinoma," *J. Clinical Pathology,* vol. 36, pp. 385-391, 1983.

[7] A. Andrion et al., "Malignant Mesothelioma of the Pleura: Inter Observer Variability," *J. Clinical Pathology,*, vol. 48, pp. 856-860, 1995.

[8] A. Young, R. Hobbs, and D. Kerr, eds., *ABC of Colorectal Cancer,* second ed. Wiley-Blackwell, 2011.

[9] J. Scholefield et al., eds., *Challenges in Colorectal Cancer,* second ed. Wiley-Blackwell, 2006.

[10] P.W. Hamilton et al., "Automated Location of Dysplastic Fields in Colorectal Histology Using Image Texture Analysis," *J. Pathology,* vol. 182, pp. 68-75, 1997.

[11] S.J. Keenan et al., "An Automated Machine Vision System for the Histological Grading of Cervical Intra Epithelial Neoplasia," *J. Pathology,* vol. 192, pp. 351-362, 2000.

[12] L. Shafarenko, M. Petrou, and J. Kittler, "Histogram Based Segmentation in a Perceptually Uniform Color Space," *IEEE Trans. Image Processing,* vol. 7, no. 9, pp. 1354-1358, Sept. 1998.

[13] P. Scheunders, "A Genetic C-Means Clustering Algorithm Applied to the Color Image Quantization," *J. Pattern Recognition,* vol. 30, pp. 859-866, 1997.

[14] T.Q. Chen and Y. Lu, "Color Image Segmentation: An Innovative Approach," *J. Pattern Recognition,* vol. 35, pp. 395-405, 2002.

[15] E. Littmann and H. Ritter, "Adaptive Color Segmentation: A Comparison of Neural and Statistical Methods," *IEEE Trans. Neural Networks,* vol. 8, no. 1, pp. 175-185, Jan. 1997.

[16] H.D. Cheng, X.H. Jiang, and J. Wang, "Color Image Segmentation Based on Homogram Thresholding and Region Merging," *J. Pattern Recognition,* vol. 35, pp. 373-393, 2002.

[17] D. Panjwani and G. Healey, "Markov Random Field Models for Unsupervised Segmentation of Textured Color Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 10, pp. 939-954, Oct. 1995.

[18] F.Y. Shih and S. Cheng, "Automatic Seeded Region Growing for Color Image Segmentation," *Image and Vision Computing,* vol. 23, pp. 877-886, 2005.

[19] S. Vicente, V. Kolmogorov, and C. Rother, "Graph Cut Based Image Segmentation with Connectivity Priors," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2008.

[20] Y. Boykov and G.F. Lea, "Graph Cuts and Effcient N-D Image Segmentation," *Int'l J. Computer Vision,* vol. 70, no. 2, pp. 109-131, 2006.

[21] H. Lu, L.P. Nolte, and M. Reyes, "Interest Points Localization for Brain Image Using Landmark-Annotated Atlas," *Int'l J. Imaging Systems and Technology,* vol. 22, no. 2, pp. 145-152, 2012.

[22] S. Bauer et al., "Integrated Segmentation of Brain Tumor Images for Radiotherapy and Neurosurgery," *Int'l J. Imaging Systems and Technology,* vol. 23, no. 1, pp. 59-63, 2013.

[23] C. Demir, S.H. Gultekin, and B. Yener, "Learning the Topological Properties of Brain Tumors," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 2, no. 3, pp. 262-270, July-Sept. 2005.

[24] M.A. Iftikhar et al., "Brain MRI Denoising and Segmentation Based on Improved Adaptive Non-Local Means," *Int'l J. Imaging Systems and Technology,* vol. 23, no. 3, pp. 235-248, 2013.

[25] M.A. Iftikhar, S. Rathore, and A. Jalil, "Parameter Optimization for Non-Local De-Noising Using Elite GA," *Proc. Int'l Multitopic Conf.,* pp. 194-199, 2012.

[26] B. Sahiner, "Classification of Mass and Normal Breast Tissue: A Convolution Neural Network Classifier with Spatial Domain and Texture Images," *IEEE Trans. Medical Imaging,* vol. 15, no. 5, pp. 598-610, Oct. 1996.

[27] H. Kobatake, Y. Yoshinaga, and M. Murakami, "Automatic Detection of Malignant Tumors on Mammogram," *Proc. IEEE Int'l Conf. Image Processing,* pp. 407-410, 1994.

[28] D. Wei et al., "Multiresolution Texture Analysis for Classification of Mass and Normal Breast Tissue on Digital Mammograms," *Proc. SPIE,* vol. 2434, pp. 606-611, 1995.

[29] J.A. Berger et al., "Jointly Analyzing Gene Expression and Copy Number Data in Breast Cancer Using Data Reduction Models," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 3, no. 1, pp. 2-16, Jan.-Mar. 2006.

[30] Y. Yuan et al., "A Sparse Regulatory Network of Copy-Number Driven Gene Expression Reveals Putative Breast Cancer Oncogenes," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 9, no. 4, pp. 947-954, July/Aug. 2012.

[31] F. Hallouche et al., "Image Processing for Cell Cycle Analysis and Discrimination in Metastatic Variant Cell Lines of B16 Murine Melanoma," *Pathobiology,* vol. 60, pp. 76-81, 1992.

[32] D.A. Diamond et al., "Computerized Image Analysis of Nuclear Shape as a Prognostic Factor for Prostatic Cancer," *Prostate,* vol. 3, pp. 321-332, 1982.

[33] D.E. Pitts et al., "Texture Analysis of Digitized Prostate Pathologic Cross-Section," *Proc. SPIE: Medical Imaging: Image Processing,* vol. 1898, pp. 465-470, 1993.

[34] M.F. McNitt-Gray, H.K. Huang, and J.W. Sayre, "Feature Selection in the Pattern Classification Problem of Digital Chest Radiograph Segmentation," *IEEE Trans. Medical Imaging,* vol. 14, no. 3, pp. 537-547, Sept. 1995.

[35] A.N. Esgiar et al., "Microscopic Image Analysis for Quantitative Measurement and Feature Identification of Normal and Cancerous Colonic Mucosa," *IEEE Trans. Information Technology in Biomedicine,* vol. 2, no. 3, pp. 197-203, Sept. 1998.

[36] A.N. Esgiar et al., "Automated Feature Extraction and Identification of Colon Carcinoma," *Analytical and Quantitative Cytology and Histology,* vol. 20, pp. 297-301, 1998.

[37] A.N. Esgiar et al., "Texture Descriptions and Classification for Pathological Analysis of Cancerous Colonic Mucosa," *Proc. Int'l Conf. Image Processing and Its Applications,* pp. 335-338, 1999.

[38] A.N. Esgiar et al., "Fractal Analysis in the Detection of Colonic Cancer Images," *IEEE Trans. Information Technology in Biomedicine,* vol. 6, no. 1, pp. 54-58, Mar. 2002.

[39] H. Kalkan, M.N.R. Duin, and M. Loog, "Automated Classification of Local Patches in Colon Histopathology," *Proc. 21st Int'l Conf. Pattern Recognition,* pp. 61-64, 2012.

[40] L. Jiao et al., "Colon Cancer Detection Using Whole Slide Histopathological Images," *Proc. World Congress on Medical Physics and Biomedical Eng.,* pp. 1283-1286, 2013.

[41] F. Ng et al., "Assessment of Primary Colorectal Cancer Heterogeneity by Using Whole-Tumor Texture Analysis: Contrast-Enhanced CT Texture As a Biomarker of 5-Year Survival," *Radiology,* vol. 266, no. 1, pp. 177-184, 2013.

[42] S. Chen, R.M. Haralick, and I. Phillips, "Extraction of Text Lines and Text Blocks on Document Images Based on Statistical Modeling," *Int'l J. Imaging Systems and Technology,* vol. 7, pp. 335-343, 1996.

[43] C. Meurie et al., "Morphological Hierarchical Segmentation and Color Spaces," *Int'l J. Imaging Systems and Technology,* vol. 20, no. 2, pp. 167-178, 2010.

[44] A.B. Tosun et al., "Object-Oriented Texture Analysis for the Unsupervised Segmentation of Biopsy Images," *J. Pattern Recognition,* vol. 42, pp. 1104-1112, 2009.

[45] A.B. Tosun, C. Sokmensuer, and C.G. Demir, "Unsupervised Tissue Image Segmentation through Object-Oriented Texture," *Proc. 20th Int'l Conf. Pattern Recognition,* pp. 2516-2519, 2010.

[46] C.G. Demir et al., "Automatic Segmentation of Colon Glands Using Object-Graphs," *Medical Image Analysis,* vol. 14, pp. 1-12, 2010.

[47] A.B. Tosun and C.G. Demir, "Graph Run-Length Matrices for Histopathological Image Segmentation," *IEEE Trans. Medical Imaging,* vol. 30, no. 3, pp. 721-732, Mar. 2011.

[48] A.C. Simsek et al., "Multilevel Segmentation of Histopathological Images Using Cooccurrence of Tissue Object," *IEEE Trans. Biomedical Eng.,* vol. 59, no. 6, pp. 1681-1690, June 2012.

[49] S. Rathore, M. Hussain, and A. Khan, "A Novel Approach for Colon Biopsy Image Segmentation," *Proc. Complex Medical Eng. Conf.,* pp. 134-139, 2013.

[50] D. Altunbay et al., "Color Graphs for Automated Cancer Diagnosis and Grading," *IEEE Trans. Biomedical Eng.,* vol. 57, no. 3, pp. 665-674, Mar. 2010.

[51] E. Ozdemir, C. Sokmensuer, and C.G. Demir, "A Resampling-Based Markovian Model for Automated Colon Cancer Diagnosis," *IEEE Trans. Biomedical Eng.,* vol. 59, no. 1, pp. 281-289, Jan. 2012.

[52] E. Ozdemir and C.G. Demir, "A Hybrid Classification Model for Digital Pathology Using Structural and Statistical Pattern Recognition," *IEEE Trans. Medical Imaging,* vol. 32, no. 2, pp. 474-483, Feb. 2013.

[53] K.M. Rajpoot and N.M. Rajpoot, "Wavelet Based Segmentation of Hyperspectral Colon Tissue Imagery," *Proc. Int'l Multitopic Conf.,* 2003.

[54] K.M. Rajpoot and N.M. Rajpoot, "SVM Optimization for Hyperspectral Colon Tissue Cell Classification," *Proc. Medical Image Computing and Computer Assisted Intervention Conf. (MICCAI '04),* pp. 829-837, 2004.

[55] K. Masood et al., "Co-Occurrence and Morphological Analysis for Colon Tissue Biopsy Classification," *Proc. Fourth Int'l Workshop Frontiers of Information Technology,* 2006.

[56] K. Masood and N. Rajpoot, "Texture Based Classification of Hyperspectral Colon Biopsy Samples Using CLBP," *Proc. IEEE Int'l Symp. Biomedical Imaging: From Nano to Macro,* pp. 1011-1014, 2009.

[57] M. Maggioni et al., "Hyperspectral Microscopic Analysis of Noral, Benign ADN Carcinoma Microarray Tissue Sections," *Proc. SPIE Optical Biopsy IV, 609101,* vol. 6091, 2006.

[58] A. Chaddad et al., "Improving of Colon Cancer Cells Detection Based on Haralick's Features on Segmented Histopathological Images," *Proc. Int'l Conf. Computer Applications and Industrial Electronics,* pp. 87-90, 2011.

[59] H. Akbari et al., "Detection of Cancer Metastasis Using a Novel Macroscopic Hyperspectral Method," *Proc. SPIE,* 2012.

[60] U. Alon, N. Barkai, and D.A. Notterman, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Science USA,* vol. 96, pp. 6745-6750, 1999.

[61] M. Grade, P. Hörmann, and S. Becker, "Gene Expression Profiling Reveals a Massive, Aneuploidy-Dependent Transcriptional Deregulation and Distinct Differences between Lymph Node-Negative and Lymph Node-Positive Colon Carcinomas," *Cancer Research,* vol. 67, pp. 41-56, 2007.

[62] S. Yajima et al., "Expression Profiling of Fecal Colonocytes for RNA-Based Screening of Colorectal Cancer," *Int'l J. Oncology,* vol. 31, pp. 1029-1037, 2007.

[63] K. Kim, U. Park, and J. Wang, "Gene Profiling of Colonic Serrated Adenomas by Using Oligonucleotide Microarray," *Int'l J. Colorectal Diseases,* vol. 23, pp. 569-580, 2008.

[64] E.T. Venkatesh, P. Thangaraj, and S. Chitra, "An Improved Neural Approach for Malignant and Normal Colon Tissue Classification from Oligonucleotide Arrays," *European J. Scientific Research,* vol. 54, pp. 159-164, 2011.

[65] A. Kulkarni et al., "Colon Cancer Prediction with Genetics Profiles Using Evolutionary Techniques," *Expert Systems with Applications,* vol. 38, pp. 2752-2757, 2011.

[66] K. Lee et al., "Classification of Bioinformatics Data Set Using Finite Impulse Response Extreme Learning Machine for Cancer Diagnosis," *Neural Computing and Applications,* vol. 22, pp. 457-468, 2013.

[67] M. Tong et al., "An Ensemble of SVM Classifiers Based on Gene Pairs," *Computers in Biology and Medicine,* vol. 43, pp. 729-737, 2013.

[68] S. Backert et al., "Differential Gene Expression in Colon Carcinoma Cells and Tissues Detected with a cDNA Array," *Int'l J. Cancer,* vol. 82, pp. 868-874, 1999.

[69] M.J. Duffy et al., "DNA Microarray-Based Gene Expression Profiling in Cancer: Aiding Cancer Diagnosis, Assessing Prognosis and Predicting Response to Therapy," *Current Pharmacogenomics,* vol. 3, pp. 289-304, 2005.

[70] M. Bianchini, E. Levy, and C. Zucchini, "Comparative Study of Gene Expression by cDNA Microarray in Human Colorectal Cancer Tissues and Normal Mucosa," *Int'l J. Oncology,* vol. 29, pp. 83-94, 2006.

[71] L. Li et al., "Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method," *Bioinformatics,* vol. 17, no. 12, pp. 1131-1142, 2001.

[72] Z. Chen and J. Li, "A Multiple Kernel Support Vector Machine Scheme for Simultaneous Feature Selection and Rule-Based Classification," *Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining,* pp. 441-448, 2007.

[73] H.S. Shon et al., "Gene Expression Data Classification Using Discrete Wavelet Transform," *Proc. Conf. Bioinformatics and Computational Biology (BioComp '09),* pp. 204-208, 2009.

[74] X. Li et al., "Detection of Colon Cancer by Laser Induced Fluorescence and Raman Spectroscopy," *Proc. IEEE Eng. in Medicine and Biology Ann. Conf.,* pp. 6961-6964, 2005.

[75] R.M. Cothren et al., "Gastrointestinal Tissue Diagnosis by Laser-Induced Fluorescence Spectroscopy at Endoscopy," *Gastrointestinal Endoscopy,* vol. 36, pp. 105-111, 1990.

[76] N.N. Boustany, R. Manoharan, and M.S. Feld, "Analysis of Normal and Diseased Colon Mucsa Using Ultraviolet Resonance Raman Spectroscopy." *Proc. SPIE,* vol. 2679, pp. 66-70, 1996.

[77] A. Molckovsky et al., "Diagnostic Potential of Near-Infrared Raman Spectroscopy in the Colon: Differentiating Adenomatous from Hyperplastic Polyps," *Gastrointestinal Endoscopy,* vol. 57, pp. 396-402, 2003.

[78] T.D. Wang et al., "Fluorescence Endoscopic Imaging of Human Colonic Adenormas," *Gastrointestinal Endoscopy,* vol. 111, pp. 1182-1191, 1996.

[79] R. Manoharan, M.S. Feld, and G. Zonios, "Laser-Induced Fluorescence Spectroscopy of Colonic Dysplasia: Prospect for Optical Histological Analysis," *Proc. SPIE, Biomedical Optics,* vol. 2388, pp. 417-421, 1995.

[80] T.D. Wang et al., "Laser-Induced Fluorescence Endoscopic Imaging for Detection of Colonic Dysplasia," *Proc. SPIE, Biomedical Optics,* vol. 2390, pp. 84-88, 1995.

[81] Z. Huang, "Laser-Induced Auto Fluorescence Microscopy of Normal and Tumor Human Colonic Tissue," *Int'l J. Oncology,* vol. 24, pp. 59-63, 2004.

[82] R.M. Cothren et al., "Detection of Dysplasia at Colon Scopy Using Laser-Induced Fluorescence: A Blinded Study," *Gastrointestinal Endoscopy,* vol. 44, pp. 168-176, 1996.

[83] R. Manoharan, Y. Wang, and M. Feld, "Ultraviolet Resonance Raman Spectroscopy for Detection of Colon Cancer," *Laser in the Life Sciences,* vol. 6, pp. 217-227, 1995.

[84] A.G. Todman, R.N.G. Naguib, and M.K. Bennet, "Visual Characteristics of Colon Images." *Proc. Medical Image Understanding and Analysis Conf.,* pp. 161-164, 2000.

[85] A.G. Todman, R.N.G. Naguib, and M.K. Bennett, "Orientational Coherence Metrics: Classification of Colon Images Based on Human from Perception," *Proc. IEEE Canadian Conf. Electrical and Computer Eng. (CCECE '01),* pp. 1379-1385, 2001.

[86] C. Demir and B. Yener, *Automated Cancer Diagnosis Based on Histopathological Images: A Systematic Survey,* Rensselaer Polytechnic Inst., Dept. of Computer Science, 2009.

[87] National Cancer Institute, "Stages of Colon Cancer," http://www.cancer.gov/cancertopics/pdq/treatment/colon/Patient/page2, June 2013.

[88] M. Fox, "Stages of Colon Adenocarcinoma," http://www.livestrong.com/article/162178-stages-of-colon-adenocarcinoma/, Jan. 2013.

[89] D. Myers, "Colon Cancer Stages: Basics of Each Colon Cancer Stage," http://coloncancer.about.com/od/stagesandsurvivalrate1/a/ColonCancerStag.htm, Jan. 2013.

[90] S. Rathore et al., "Texture Analysis for Liver Segmentation and Classification: A Survey," *Proc. Frontiers of Information Technology,* pp. 121-126, 2011.

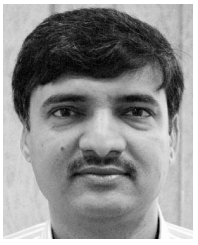
[91] R.C. Gonzalez and R.E. Woods, *Digital Image Processing.* Prentice Hall, 2002.

[92] K. Fukunaga and L. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," *IEEE Trans. Information Theory,* vol. 21, no. 1, pp. 32-40, Jan. 1975.
[93] C.G. Demir, "Mathematical Modeling of the Malignancy of Cancer Using Graph Evolution," *Math. Biosciences,* vol. 209, no. 2, pp. 514-527, 2007.
[94] H.S. Wu et al., "Segmentation of Intestinal Gland Images with Iterative Region Growing," *J. Microscopy,* vol. 220, no. 3, pp. 190-204, 2005.
[95] S. Naik et al., "Gland Segmentation and Gleason Grading of Prostate Histology by Integrating Low-High-Level and Domain Specific Information," *Proc. Second Workshop Microscopic Image Analysis with Applications in Biology,* 2007.
[96] M.M. Galloway, "Texture Analysis Using Gray Level Run Lengths," *Computer Graphics and Image Processing,* vol. 4, no. 2, pp. 172-179, 1975.
[97] G. Davis, M. Maggioni, and R. Coifman, "Spectral Spatial Analysis of Colon Carcinoma," *J. Modern Pathology,* vol. 16, pp. 320-332, 2003.
[98] T. Marks, *Principal Component Analysis, Lecture Notes,* Cognitive Science Department, Univ. of California at San Diego, 2001.
[99] S. Balakrishnama and A. Ganapathiraju, "Linear Discriminant Analysis—A Brief Tutorial," Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State Univ., 1998.
[100] T. Chan and L. Vese, "Active Contours without Edges," *IEEE Trans. Image Processing,* vol. 10, no. 2, pp. 266-277, Feb. 2001.
[101] Z. Man et al., "A New Robust Training Algorithm for a Class of Single-Hidden Layer Feedforward Neural Networks," *Neurocomputing,* vol. 74, no. 16, pp. 2491-2501, 2011.
[102] Nikon, "Nikon Coolscope Digital Microscope," http://www.microscopyu.com/tutorials/java/coolscope/index.html, 2013.
[103] R. Sanz-Pamplona et al., "Clinical Value of Prognosis Gene Expression Signatures in Colorectal Cancer: A Systematic Review," *PLoS ONE,* vol. 7, no. 11, article e48877, 2012.
[104] "Colon Cancer Data Set Kent Ridge," http://datam.i2r.a-star.edu.sg/datasets/krbd/ColonTumor/ColonTumor.html, 2013.
[105] BioGPS, "Dataset: Stage II and Stage III Colorectal Cancer," http://biogps.org/dataset/1352/stage-ii-and-stage-iii-colorectal-cancer/, 2013.

Saima Rathore received the BS degree in software engineering from Fatima Jinnah Women University, Rawalpindi, Pakistan, in 2006 and the MS degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2008. Currently, she is working toward the PhD degree at the Pakistan Institute of Engineering and Applied Sciences, Islamabad. She possesses more than eight years of research and teaching experience at the university level. Her research interests include medical image analysis, segmentation, classification and evolutionary algorithms.

Mutawarra Hussain received the MSc degree in computer science and the MS degree in nuclear engineering from Quaid-i-Azam University, Islamabad, Pakistan, in 1979 and 1981, respectively, and the PhD degree in the field of medical image analysis from the School of Computer Science, University of Birmingham, United Kingdom, in 2000. He worked as a postdoctoral researcher in the field of medical image analysis at the School of Computer Science, University of Birmingham, United Kingdom, in 2008. He has more than 30 years of research experience and is currently working as the dean of research at the Pakistan Institute of Engineering and Applied Sciences, Islamabad. His research interests include medical image analysis, evolutionary algorithms, image segmentation, and classification.

Ahmad Ali received the BS degree in computer sciences (Hons) from the University of Engineering and Technology, Lahore, Pakistan, in 2002 and the MS degree in systems engineering from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, in 2005. Currently, he is working toward the PhD degree from PIEAS. He joined the Centers of Excellence in Science and Applied Technologies, Islamabad, in 2005. His research interests include image processing and computer vision. He won the National Fellowship Competition Award (2003-2005) and received the IT and Telecom Endowment Fund Scholarship Award for PhD in 2010.



Asifullah Khan received the MSc degree in physics from the University of Peshawar, Pakistan, in 1996, the MS degree in nuclear engineering from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan, in 1998, and the MS and PhD degrees in computer systems engineering from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIK Institute), Topi, Pakistan, in 2003 and 2006, respectively. He has completed two years of postdoctoral research at the Signal and Image Processing Lab, Department of Mechatronics, Gwangju Institute of Science and Technology, Korea. He has more than 15 years of research experience and is working as an associate professor in the Department of Computer and Information Sciences at PIEAS. His research interests include digital watermarking, pattern recognition, image processing, and evolutionary algorithms.