Automated labelling of cancer textures in colorectal histopathology slides using quasi-supervised learning

Devrim Onder [a,*], Sulen Sarioglu [b], Bilge Karacali [a]

[a] Electrical and Electronics Engineering Department, Izmir Institute of Technology, Urla, Izmir 35430, Turkey
[b] Department of Pathology, Faculty of Medicine, Dokuz Eylul University, Inciralti, Izmir 35340, Turkey

ARTICLE INFO

ABSTRACT

Quasi-supervised learning is a statistical learning algorithm that contrasts two datasets by computing estimate for the posterior probability of each sample in either dataset. This method has not been applied to histopathological images before. The purpose of this study is to evaluate the performance of the method to identify colorectal tissues with or without adenocarcinoma. Light microscopic digital images from histopathological sections were obtained from 30 colorectal radical surgery materials including adenocarcinoma and non-neoplastic regions. The texture features were extracted by using local histograms and co-occurrence matrices. The quasi-supervised learning algorithm operates on two datasets, one containing samples of normal tissues labelled only indirectly, and the other containing an unlabeled collection of samples of both normal and cancer tissues. As such, the algorithm eliminates the need for manually labelled samples of normal and cancer tissues for conventional supervised learning and significantly reduces the expert intervention. Several texture feature vector datasets corresponding to different extraction parameters were tested within the proposed framework. The Independent Component Analysis dimensionality reduction approach was also identified as the one improving the labelling performance evaluated in this series. In this series, the proposed method was applied to the dataset of 22,080 vectors with reduced dimensionality 119 from 132. Regions containing cancer tissue could be identified accurately having false and true positive rates up to 19% and 88% respectively without using manually labelled ground-truth datasets in a quasi-supervised strategy. The resulting labelling performances were compared to that of a conventional powerful supervised classifier using manually labelled ground-truth data. The supervised classifier results were calculated as 3.5% and 95% for the same case. The results in this series in comparison with the benchmark classifier, suggest that quasi-supervised image texture labelling may be a useful method in the analysis and classification of pathological slides but further study is required to improve the results.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Computerized analysis of histopathology slides has been a very attractive research topic with the recent advances in computational power. Meanwhile, improvements in image analysis and machine learning techniques allowed researchers to address the ultimate goal of supporting pathologists in diagnosis, disease detection and grading. Increasing number of complex patterns that need to be checked by pathologists and rapidly growing histopathology slide databases keep the subject of automated quantitative analysis of histopathology slides indispensable today. There were many studies conducted for automated detection of regions having the characteristics of a specific disease. The diseases taken into

consideration in this respect include colorectal dysplasia (Hamilton et al., 1997), breast lesions (Sahiner et al., 1996; Dundar et al., 2010, 2011), renal cell carcinoma (Waheed et al., 2007), cervical (Hallouche et al., 1992), prostate (Diamond et al., 1982; Pitts et al., 1993; Doyle et al., 2006, 2007; Huang and Lee, 2009), oral cancers (Muthu et al., 2012; Mookiah et al., 2011) and colon cancers (Hamilton et al., 1987; Nasser Esgiar et al., 1998; Rajpoot and Rajpoot, 2004; Masood et al., 2006; Filippas et al., 2003; Nwoye et al., 2006).

The classification accuracy in colorectal samples was 83% in dysplasia (Hamilton et al., 1997) and 88% in inflammatory colorectal diseases (Ficsor et al., 2008). The accuracy acquired in identification of colorectal cancers was 89–100% in different series (Nasser Esgiar et al., 1998; Rajpoot and Rajpoot, 2004; Masood et al., 2006; Filippas et al., 2003; Nwoye et al., 2006) by using various methods.

The common machine learning approach in all of these studies was the use of a supervised classification procedure to label the tissue regions of interest. In these supervised classification

* Corresponding author. Tel.: +90 232 750 65 10; fax: +90 232 750 65 99.
  E-mail addresses: devrimonder@gmail.com, devrimonder@iyte.edu.tr
(D. Onder).

studies, manually labelled training data were needed for supervised classifier training. Among the classification strategies put to the task both linear discriminate function and K-Nearest-Neighbour non-parametric classifiers were separately used to identify cancerous colonic mucosa (Nasser Esgiar et al., 1998) while in other series, a support vector machine (SVM) classifier was used to carry out the discrimination between normal and malignant colon tissue cells (Rajpoot and Rajpoot, 2004; Masood et al., 2006). Filippas et al. (2003) focused on the identification of normal and cancerous colonic mucosa using a genetic algorithm. Doyle et al. performed studies for automated detection of prostatic adenocarcinoma and for prostate cancer grading using Adaboost, Decision Trees and SVM classifiers (Doyle et al., 2006, 2007). There are other methods applied to the different types of carcinomas. In renal carcinomas, computer-aided histopathological classification using a multi-class Bayesian decision rule that assumes multivariate Gaussian distributions for the feature vectors was also used (Waheed et al., 2007). Nwoye et al. used a fuzzy neural network classifier to detect adenomas and adenocarcinomas in colorectal tissue slides (Waheed et al., 2007; Nwoye et al., 2006). Krishnan et al. found the best method among five different classifiers; Decision Tree (DT), Sugeno Fuzzy, Gaussian Mixture Model (GMM), K-Nearest-Neighbour (K-NN) and Radial Basis Probabilistic (Muthu et al., 2012) classifiers.

Besides these approaches, the Multiple Instance Learning (MIL) procedure has been proposed as a variation of supervised learning for problems with incomplete knowledge about labels of training samples (Dietterich et al., 1997). In MIL, the labels are only assigned to bags of instances. In the binary case, a bag is labelled positive if at least one instance in that bag is positive, and the bag is labelled negative if all the instances in it are negative. The goal of MIL is to classify unseen bags or instances based on the labelled bags as the training data. This classification methodology was used in several applications of the automated diagnosis of breast lesions (Dundar et al., 2010, 2011) and classification human histological images (Zhao et al., 2006).

Beyond the construction of a texture classifier using one of classifier construction alternatives, all these methods share a common ground: a manually labelled collection of regions on histopathology slides. The process of obtaining such a collection, however, can be painstaking and laborious, especially when the number of histopathology slides on which the manual labelling is carried out is large to achieve better statistical characterization. In addition, the inter-observer and intra-observer variability of pathologists interpretation of histopathology slides is a well recognized issue (Sarioglu et al., 2010) which may even effect the common supervised machine learning approach.

Quasi-supervised learning (QSL) is a statistical learning algorithm that contrasts two datasets by computing estimates for the posterior probabilities of their samples of belonging in either dataset (Karacali, 2010). The QSL method addresses an identification problem where labelled samples are available from one class only, in a reference dataset. A second, unlabelled dataset is also provided and contains a mixture of samples from both control and target classes.

In performance evaluation experiments on synthetic datasets, QSL method outperformed alternative strategies based on SVM classification and minimum spanning trees for varying dataset size, overlap, and dimensionality (Karacali, 2010). The QSL algorithm was successfully operated on a limited number of colon histopathology slide image database (Onder et al., 2010). Kokturk also studied the separation of the electroencephalography data recorded under different visual stimuli by using the QSL algorithm (Kokturk, 2011). The data used in that study contained multiple channel EEG recordings under six different visual stimuli in random successive order. Kokturk identified condition-specific EEG profiles in different comparison scenarios by using the QSL. The results revealed that the

QSL algorithm was efficient in capturing the distinction between the experimental data samples.

This article presents the first application of the quasi-supervised learning algorithm to fully automated labelling of cancer histology slides. In the quasi-supervised learning setting, a manually curated ground truth dataset required for conventional classifier-based methods is not available for learning. Instead, the proposed methodology contrasts the regions observed on histology slides of non-neoplastic tissues to those containing tumour, and automatically identifies the tumour-related regions that are highly specific to the slides of tumour tissue. As the slides of the tumour tissue also contain non-neoplastic tissue regions, separation of the tumour tissues from the surrounding tissues in the absence of previously labelled examples represents a novel labelling problem that cannot be addressed using any of the previous methods such as those based on support vector machines. In the experiments on a large collection of colorectal carcinoma (CRCa) histology slides, the proposed methodology successfully identified the tumour regions without any manual labelling for representative regions of non-neoplastic and cancer tissues. As such, it offers a viable alternative to classifier-based strategies for automated labelling of histology slides by eliminating the need for manually curating ground-truth datasets characterizing all tissue classes of interest and thereby minimizing the training effort.

2. Materials and methods

2.1. Histopathological Image Library

An image library of 230 light microscopic images was constructed to be used in the texture classification experiments. This set of digital light microscopic images was taken from heamatoxylin and eosin (H&E) stained sections of formalin fixed paraffin embedded tissue sections of non-neoplastic colorectal (NNCR) and CRCa tissues from radical colectomy or rectum resection specimens by a camera (Olympus DP70, Olympus Optical Co. Ltd., Tokyo, Japan) connected to a light microscope (Olympus BX51, Olympus Optical Co. Ltd., Tokyo, Japan) at an original magnification of 4×.

All the original images in the library have 256 level RGB planes, 4080×3720 pixel dimensions and approximately 1 μm per pixel resolution. 4×4 pixels regular grid sampling was applied to all original images to get smaller image size that yielded relatively shorter computation times in feature extraction stage. Thus, the secondary histopathological images used throughout this research had 4 μm per pixel resolution.

The images were acquired using fixed capture and illumination parameters (the microscope light exposure was manually set to 6 from a scale of 0 to 6). The image library was then divided into two groups as NNCR and unlabelled mixed; NNCR and CRCa. The rule for group assignment was as follows; when an expert observed no carcinoma region throughout an image, that image was labelled as NNCR and assigned to the first group.

For the second group, the term "mixed" is used in order to indicate that these images are composed of features associated with both NNCR and CRCa tissues. This separation of the images into two groups by a pathology expert was a very simple task compared to manual labelling of isolated colorectal carcinoma regions and it was much less operator dependent. The NNCR group had 127 images and the mixed group had 103 images in total.

2.2. The proposed framework

The proposed framework is illustrated in Fig. 1, in the form of a flow diagram combining the individual methods described in this section.
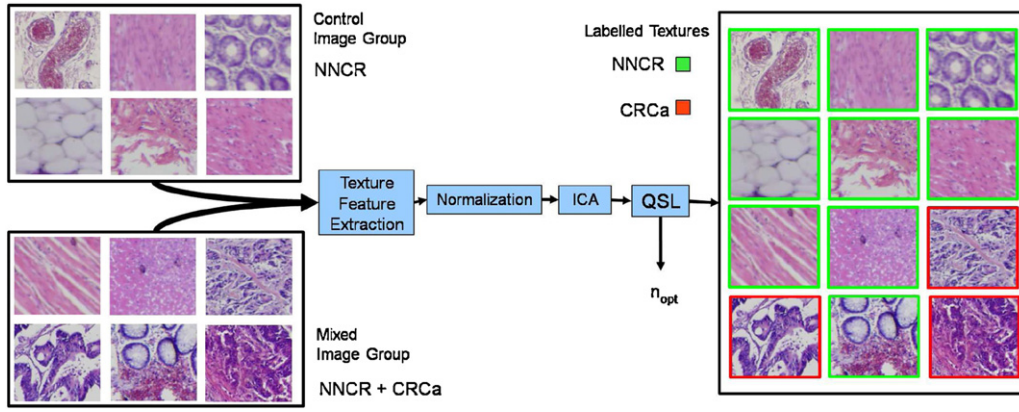
Fig. 1. Graphical abstract of the proposed QSL texture labelling framework.

The first block of the proposed framework is texture feature extraction which will be explained in the following section. In this block, the texture feature vectors are calculated using the histopathology images from the reference and the mixed image groups. Each feature vector takes the initial label of its source image group; either reference or mixed.

The blocks following the texture feature extraction are the vector normalization and the vector dimensionality reduction blocks. After the dimensionality reduction, all of the reduced feature vectors are then fed to the QSL algorithm. The QSL algorithm calculates the posterior probability $p_0(\mathbf{x})$ for each texture feature vector $\mathbf{x}$, defined as the probability of it being assigned to the reference label. At the same time, the optimum reference set size n_{opt} for these texture vector datasets is also calculated.

After $p_0(\mathbf{x})$ values are calculated for all feature vectors $\mathbf{x}$, the threshold T_{opt} is to be determined in order to set resultant texture vector labels. The feature vectors $\mathbf{x}$ that satisfy the condition $p_0(\mathbf{x}) \geq T_{opt}$ are assigned to the label reference, whereas, all other vectors are assigned to the label of contrast. The meaning of the contrasted label may vary due to the experimental setup but it is basically the contrasted characteristics between reference and mixed image groups. In our study, the reference image label corresponds to the healthy regions (NNCR) whereas contrasted label, i.e. the contrast between the mixed and the reference groups, means cancer (CRCa) or any other unknown local characteristics specific to the mixed image group.

To display the labelling results, all local image regions are framed as an overlay onto the original histopathology images by using two different colours and line styles. The green dashed lines represent textures labelled as healthy and the solid red lines represent textures labelled as cancer-related.

2.3. Texture feature vectors

This section describes the texture feature extraction strategy and presents the several different vector datasets used in this study.

2.3.1. Constitution of texture feature vectors

One of the classical approaches to texture classification is to use texture features derived from co-occurrence matrices (Haralick et al., 1973). A co-occurrence matrix is a local approximation to the joint distribution of grey level values of pixel pairs at specified distances. The entry of a co-occurrence matrix at the ith row and the jth column in an image I is calculated by:

$$M_{I,d}(i,j) = \frac{\sum_{p,q \in B_r(x)} 1\{I(p) = i, I(q) = j, \rho(p,q) = d\}}{\sum_{p,q \in B_r(x)} 1\{\rho(p,q) = d\}} \qquad (1)$$

where I is the grey level image, d is the pair-wise distance of pixels and $B_r(x)$ defines the neighbourhood of radius r around a centre coordinate x. Furthermore, p and q represent two pixels in the image and $\rho(p,q)$ represents the Euclidean distance between them. There is no direction specific relation exists between p and q while selecting these image pixels. The function (1) is defined as a binary function that takes the values 1 or 0 when its argument is true and false, respectively.

Although it is theoretically possible to calculate texture feature vectors around every pixel of an image under consideration, it is not computationally feasible. Therefore, we have assumed that the co-occurrence matrix varied smoothly across the image, and carried out feature vector computation for points on a regular grid with r pixels spacing. Furthermore, the decision made for a feature vector of a grid point was generalized to the square image region of size $r \times r$ centred at that point. The sketch for the calculation geometry is shown in Fig. 2.

Texture features that characterize the appearance of an image square are composed of the first order and the second order characteristics. The first order texture features were obtained using local histograms, whereas the second order features were obtained from co-occurrence matrices calculated in a local image region. The list of texture features used in this study is given below. The second order texture feature characteristics except (2.(n)) are known as Haralick features (Haralick et al., 1973).

1. First order features (Pratt, 1991; Gonzalez and Woods, 1992)
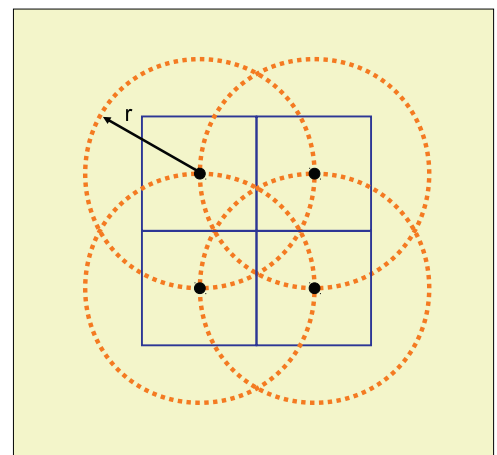   (a) Mean value of pixels,

Fig. 2. Texture feature vector calculation geometry.

(b) Variance,
(c) Skewness,
(d) Kurtosis,
(e) Entropy,
(f) Energy,
(g) Maximum value of histogram,
(h) Corresponding pixel value for maximum value of histogram.
2. Second order features
(a) Angular second moment,
(b) Contrast,
(c) Correlation,
(d) Sum of squares: variance,
(e) Inverse difference moment,
(f) Sum average,
(g) Sum entropy,
(h) Sum variance,
(i) Entropy,
(g) Difference variance,
(k) Difference entropy,
(l) Information measures of correlation 1,
(m) Information measures of correlation 2,
(n) Maximum probability (maximum co-occurrence matrix element).

For a local image square, each texture feature in the list above constitutes a component of its texture feature vector. It should be noted that for a local image square, various second order texture features can be calculated for co-occurrence matrices obtained using different pair-wise pixel distance values d. Second order feature vector components calculated using various d values can then be combined with the first order features to form a more detailed feature vector.

Another strategy to enrich the set of texture features is to use a hierarchical organization. In a hierarchical computation of texture features, a set of radius values that are multiples of r is used to define a nested set of neighbourhoods. For each neighbourhood, texture feature vector extraction is performed and these vectors are then concatenated in order to have higher dimensional resultant texture feature vector. In this strategy, (i, j)th entry of a co-occurrence matrix of a hierarchical level $h = 1, 2, \ldots, H$ is calculated by;

$$M_{I,h,d}(i,j) = \frac{\sum_{p,q\in B_{r\times h}(x)} 1\{I(p) = i, I(q) = j, \rho(p, q) = d\}}{\sum_{p,q\in B_{r\times h}(x)} 1\{\rho(p, q) = d\}}, \quad (2)$$

similar to Eq. (1). The difference is the size of neighbourhood $B_{r\times h}(x)$. Note that, $h = 1$ corresponds to non-hierarchical texture feature extraction. The idea behind hierarchical texture feature vectors is to identify small scale texture characteristics together with those present at larger scales. The geometry of hierarchical texture feature calculation for $h = \{1, 2\}$ is shown in Fig. 3.

2.3.2. Texture feature vector datasets

Throughout this study, various texture feature vector datasets were constructed in order to evaluate the corresponding texture labelling performances. A texture feature dataset is differentiated with the parameters used in feature vector extraction. Texture extraction configuration parameters used in this study are listed in groups as below;

1. Source image plane(s),
(a) Grey level plane,
(b) Lab colour planes (L, a and b).
2. Basic radius value r of texture feature calculation geometry,
3. Set of pairwise pixel distance values used together d,
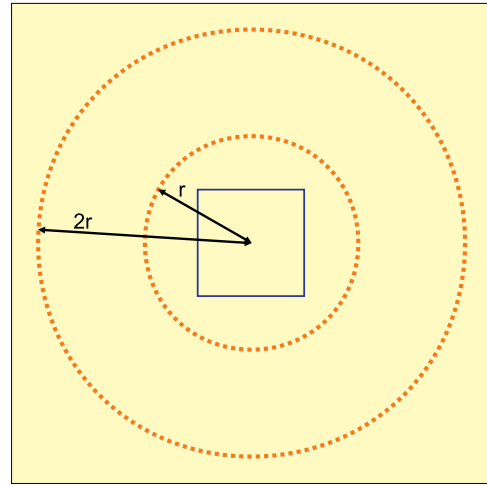4. Hierarchical feature computation.

Fig. 3. Hierarchical texture feature vector calculation geometry for $h = \{1, 2\}$.

(a) Single neighbourhood ($H = 1$),
(b) Multiple nested neighbourhoods ($H \geq 2$).

In the experiments of this study, both the grey level and the colour information obtained from the histopathological images were used in parallel. In case of colour image processing, original images were transformed into Lab image planes using the well-known RGB to Lab colour transformation referencing a white point of (255, 255, 255) in the RGB space, and each image plane was processed separately (Schwarz et al., 1987; CIE, 1986). Specifically, for an image region, texture feature vectors were calculated for each of L, a and b planes and then concatenated to produce a single texture feature vector. Therefore, a texture feature vector obtained from an Lab image had three times the dimension as the one obtained from a grey level image. In addition, uniform scalar quantization on 16 intensity levels was performed on each image plane of interest before calculating the co-occurrence matrices and the texture feature vector components to limit the number of possible image intensity pairs.

There are several limitations in selecting a radius value r for the feature calculation geometry. Firstly, the radius value selected was to match the discriminative local texture characteristics. Large radius values yield less number of feature vectors than the smaller values in total, and sustain difficulties in defining the regions of texture transition. On the other hand, smaller radius values provide relatively higher separation in texture transition regions but make the labelling problem labour intensive due to larger numbers of texture feature vectors. In addition, the smallest artifact observed in the Histopathological Image Library, the cell nucleus was 15–20 μm in diameter. As a result, a list of r values, 32, 48, 64, 128 pixels were taken as alternative texture extraction configuration parameters.

For each texture feature vector dataset, a set of pair-wise pixel distance value(s) d was arbitrarily taken into consideration with the upper boundary equals to the basic radius r.

The level of hierarchy H was either selected as 1 or 2, limiting the feature vector computations to neighbourhoods of radii r and $2r$. $H = 1$ case is also regarded as the "no hierarchy" case. The resulting texture feature vector datasets of this study are listed in Table 1 with their respective texture feature extraction configuration parameters:

2.4. Quasi-supervised learning

Quasi-supervised learning (QSL), is a machine-learning technique that uses both indirectly labelled and unlabelled data as

Table 1
Texture configuration parameter values for each texture feature vector dataset.

Texture features	Source plane(s)	r	H	d
Dataset 1	Grey level	64	2	{1, 3, 5, 9, 13, 17, 21, 41, 51, 61}
Dataset 2	Lab	64	2	{1, 3, 5, 9, 13, 17, 21, 41, 51, 61}
Dataset 3	Lab	64	1	{1, 3, 5, 9, 13, 17, 21, 41, 51, 61}
Dataset 4	Lab	48	2	{1, 3, 7, 13}
Dataset 5	Lab	64	2	{1}
Dataset 6	Lab	32	2	{1, 3, 7, 13}

training datasets and does not require manually labelled ground-truth vector datasets for recognition (Karacali, 2010). In a binary recognition setting, the algorithm is provided with two datasets: a reference dataset C_0 containing samples of one class only, and a mixed dataset C_1 containing an unlabelled collection of samples. The algorithm then exploits the asymptotic properties of Nearest-Neighbourhood classification on randomly selected reference sets (Cover and Hart, 1967). Specifically, it estimates the posterior probabilities $Pr\{C_0|x\}$ and $Pr\{C_1|x\}$ using an average number of times a point x is assigned to C_0 or C_1 with respect to a random reference set containing n points from each collection.

The posterior probability forms the basis for automatic labelling of the samples in the unlabelled dataset. Given a threshold T, the unlabelled dataset samples $\mathbf{x}$ for which $p_1(\mathbf{x}) \geq T$ are identified as samples specific to C_1, and by construction, belonging to the class not represented in C_0. Finally, the reference set size parameter n is determined by minimizing the following cost functional:

$$E(n) = 4\sum_i (p_0(\mathbf{x})p_1(\mathbf{x})) + 2n \tag{3}$$

The first term in this cost functional represents the penalty for the large class overlaps, while, the second term represents the preference for smaller n values to achieve better generalisation with Nearest-Neighbourhood classification. The reasoning and the verification of the cost functional were explained in greater detail in Karacali (2010).

In this study, the QSL algorithm was applied to the texture feature vectors corresponding to the reference and mixed library image groups and the subsequent labelling of the corresponding image regions were obtained. The procedure first computed the posterior probabilities $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ for all feature vectors in the two datasets for the optimal reference set size n. During the computation of the posterior probability for a given n, the feature vectors obtained from the same image as the vector in consideration were left outside of the analysis in order to prevent biasing the analysis results due to the possible similarity between feature vectors obtained from the same histopathology slide.

In the labelling of a feature vector $\mathbf{x}$, higher $p_0(\mathbf{x})$ value means that the vector in question is more similar to those of NNCR tissues in the reference dataset. In turn, lower $p_0(\mathbf{x})$ value means that the vector in question is different from the NNCR feature vectors and by construction associated with CRCa.

2.5. Texture labelling performance evaluation

This section describes the strategy that uses ground truth texture information in order to evaluate the labelling performance of the proposed framework.

2.5.1. Ground truth atlas data

Texture feature vector datasets and corresponding group labels as reference or unlabelled were fed to QSL algorithm and the resulting labels compared with the ground-truth vector labels in order to measure the labelling performance. To obtain ground-truth atlas data, a software tool, the *Histopathological Image Atlas Editor* (HIAE)

was developed in C++with a Graphical User Interface for the Windows operating system. The HIAE retrieves selected images from the Histopathological Image Library and allows an expert to mark the cancer regions by mouse. Each image in the library was divided into a grid of 128 μm and the labelling was manually performed for each square block using the HIAE. This data constitutes the ground-truth data used to evaluate the performance of the proposed method. The atlas data prepared by using the HIAE, overlaid on two colorectal histopathology images are shown in Fig. 6(a) and (b). In these figures, individual square regions that were marked by the expert were merged and framed as an overlay onto original histopathology images.

An important point to emphasize here is that the ground-truth data was collected and used for the purpose of evaluating the performance of the proposed histopathology slide labelling method. While such datasets are required for training conventional supervised classification methods, the quasi-supervised learning paradigm adopted here is designed explicitly to remove the need for ground-truth training datasets in learning. Hence, the ground-truth dataset was withheld from the quasi-supervised learning in the experiments.

2.5.2. Receiver operating characteristics curve

In order to assess the separation of the NNCR and CRCa tissue regions, receiver operating characteristics (ROC) curves were constructed. An ROC curve is a graphical plot of the true positive rate versus the false positive rate. The true positive rate, P_{TP}, is the probability of successful labelling of all ground-truth cancer features vectors. Similarly, the false positive rate, P_{FP}, is the probability of labelling NNCR features as cancer. In order to generate an ROC curve, we have computed P_{FP} on the reference vector dataset and P_{TP} on the unlabelled vector dataset, and plotted for varying threshold values T ranging from 0.0 to 1.0.

We have evaluated the labelling performance corresponding to an ROC curve via two separate strategies:

1. The area under the ROC curve: The area under the ROC curve is a performance measure in which the larger area means a better separation of the NNCR and CRCa tissue regions. In the ideal case, the area under the ROC curve would be equal to 1.0.
2. The optimum recognition point on the ROC curve: The optimum threshold value T_{opt} is selected on the neck point of the continuous ROC curve where its slope equals to 1.0. In the ideal case, the ROC curve would be the unit step function and the optimum recognition point on this curve would be at the (0.0, 1.0) point. This ideal point means that there is no false alarm with full true detection. After T_{opt} is determined, the final labelling of the images was carried out using this threshold value and the corresponding P_{FP} and P_{TP} values were registered as the performance measures of that labelling experiment.

When comparing the results of any two experiments, the one with the larger area under the ROC curve was identified as achieving a better identification. If two experiments had very close ROC curve areas, then the one with the optimum recognition point (P_{FP}, P_{TP}) closer to ideal point (0.0, 1.0) was accepted as more successful identification.

2.6. Comparative classification

In order to obtain an independent evaluation of the labelling performance, we have also used a support vector machine (SVM) classifier trained on the ground-truth label data that was withheld from the quasi-supervised labelling strategy (Cortes and Vapnik, 1995; Vapnik, 1998; Burges, 1998). In the SVM classification

experiments, we used a third party implementation, SVM^{light} (http://svmlight.joachims.org).

In order to perform the SVM classification experiments, we constructed the NNCR and CRCa vector groups using ground-truth atlas information. Group vectors took labels +1 and −1, regarding to the condition if they belong to NNCR group or not. An SVM classifier training vector set was then constructed by randomly selecting 90% of these vectors and, a control vector set was constructed by the remaining 10%. The training feature vector set was used to obtain the classifier model, and this classifier model was operated on control feature vector set. The SVM classifier output, obtained for the vectors in the control set, was compared with their labels in the Histopathological Image Atlas to calculate the resultant classification performance.

Since a relatively smaller number of texture vectors were used in the control set of an individual SVM experiment, the performance measures obtained were not expected to represent the actual classifier performance well. To address this issue, multiple independent SVM classification experiments (40 in our study) were performed for a specific texture vector dataset, and the resultant classification performance were used to determine the corresponding statistics, the mean and the variance. The mean and the variance statistics determined the performance of the classifier on a specific texture vector dataset.

In the SVM classifier construction, we have used a Gaussian radial basis function kernel with

$$K(x_i, x_j) = \exp{-\frac{\|x_i, x_j\|^2}{2\sigma^2}} \tag{4}$$

where the scale parameter σ was determined by minimizing the number of support vectors in the training phase via a line search. In order to take into account the non-separable cases, the Lagrange multipliers of the quadratic optimization were bound from above by 100.0 during training, producing a soft-margin classification. The recognition performance of the samples in the control dataset was carried out by thresholding the classifier underlying function

$$h(x) = \sum_i y_i \beta_i K(x, x_j) + b \tag{5}$$

by a threshold $T\varepsilon(-\infty, +\infty)$, y_i being +1 or −1 based on whether x_i belongs to the NNCR or CRCa groups respectively, and β_i and b were obtained by training the classifier. The control dataset samples for which $h(x) \geq T$ were then recognized as NNCR.

2.7. Vector dimensionality reduction

Vector dimensionality reduction is a mathematical transformation to represent a vector dataset in a relatively lower dimension. The dimensionality reduction is usually performed to improve the classification performance and the computation time. In our study, we have considered several vector dimensionality reduction methods, namely, Principal Component Analysis (PCA) (Jolliffe, 2002), Isomap (Tenenbaum et al., 2000) and a basic feature subset selection using a separation criteria as well as Independent Component Analysis (ICA). ICA is a vector dimensional reduction approach that seeks to express a multivariate distribution as a linear combination of statistically independent non-Gaussian random variables (Comon, 1994; Hyvärinen et al., 2001) (http://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml). The statistically independent variables are called as the independent components of the observed multi-dimensional signal. The ICA method is widely used in the areas of signal source separation and feature extraction.
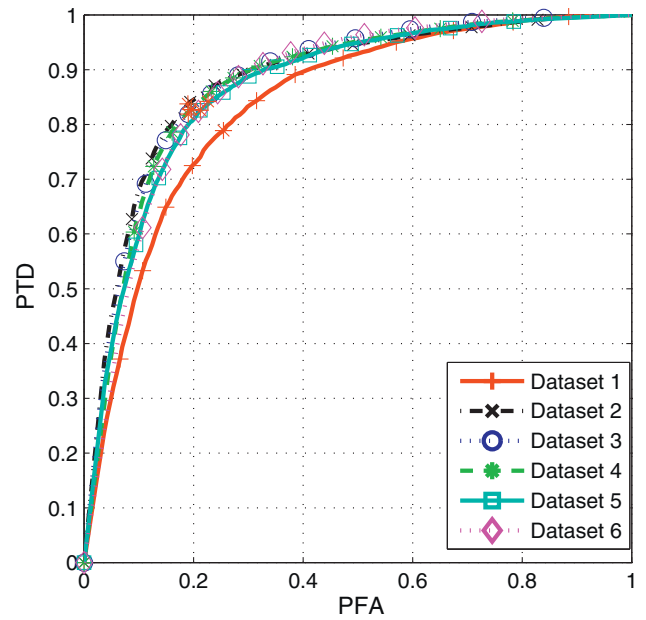

Fig. 4. ROC curves (Datasets 1–6).

3. Results

This section presents the results of the experimental execution; the resultant labelling performances are listed both for original and dimensionally reduced vector datasets. In addition, performances of the SVM supervised classifier operated on the original texture feature vector datasets were presented for comparison to an idealized scenario.

3.1. Texture labelling performances

The ROC curves obtained by the QSL algorithm operated on Datasets 1–6 are given in Fig. 4. According to the performance comparison rules described in Section 2.5 it was observed that the best recognition performance was acquired with texture vector Dataset 2. It was also observed that, if an experiment had better optimum recognition point than another one, it had also larger ROC curve area.

3.2. Comparative classification case

The QSL algorithm labelling performances and the SVM classifier performances for texture feature Datasets 1–6 are presented in Table 2. Note that, while labelling performance measure values are given directly for the QSL experiments, the classification performance statistics are given for the SVM experiments as explained before. It can be observed that the SVM classifier performance P_{FP} values are distributed around relatively lower values compared to that of QSL. On the other hand, P_{TP} values are much close for both

Table 2
Performances of the QSL method and the SVM classifier.

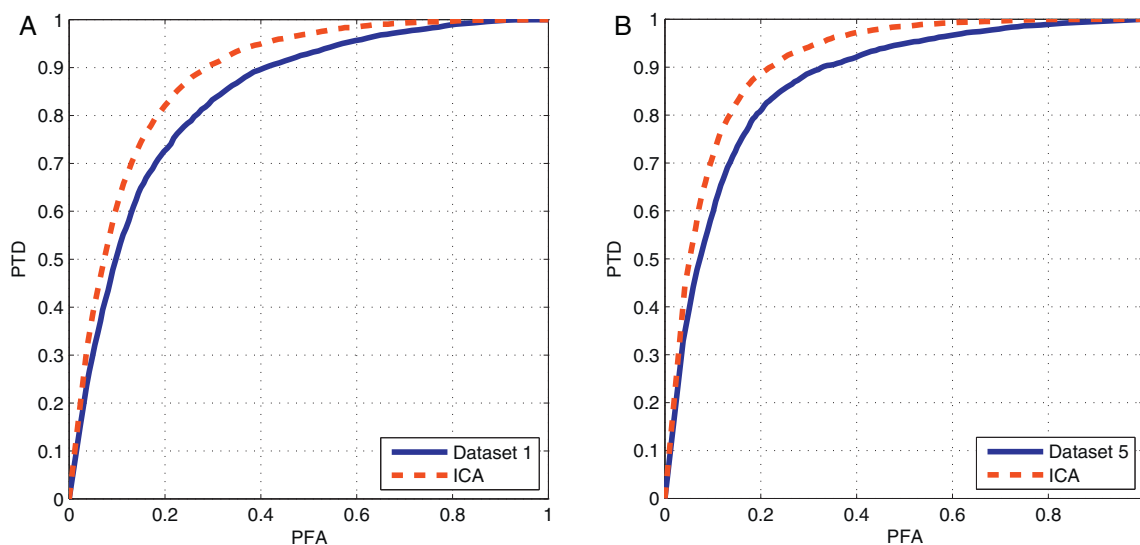| Texture features	QSL		SVM	
	(P_{FP}, P_{TP})	ROC area	$P_{FP}(\mu \mp \sigma)$	$P_{TP}(\mu \mp \sigma)$
Dataset 1	(0.25, 0.79)	0.84	0.15 ∓ 0.01	0.87 ∓ 0.01
Dataset 2	(0.19, 0.84)	0.88	0.03 ∓ 0.00	0.96 ∓ 0.00
Dataset 3	(0.19, 0.82)	0.88	0.03 ∓ 0.00	0.94 ∓ 0.00
Dataset 4	(0.20, 0.83)	0.88	0.04 ∓ 0.02	0.86 ∓ 0.26
Dataset 5	(0.21, 0.83)	0.87	0.04 ∓ 0.03	0.84 ∓ 0.40
Dataset 6	(0.23, 0.84)	0.87	0.02 ∓ 0.00	0.96 ∓ 0.00

Fig. 5. Two ROC curve pairs for the baseline and the ICA applied vector datasets. The baseline and the reduced vector dimensions for (a) Dataset 1 are 296 and 279 and (b) Dataset 5 are 132 and 119.

methods. Note, however, that as the SVM classifier was trained on the atlas data, it represents an upper bound to the labelling performance when there is no ground-truth data to be used in learning. The results in Table 2 indicate that the QSL strategy attains a labelling performance that is close to this upper bound without the benefit of a ground-truth learning dataset.

3.3. Dimensionality reduction by ICA

The FastICA algorithm (http://research.ics.tkk.fi/ica/fastica) is a popular ICA method that uses a fixed point iteration scheme to maximise non-Gaussianity of the sources. In independent experiments, FastICA has been found to be 10–100 times faster than conventional gradient descent based ICA methods.

In the experiments of this study, the original texture feature vector datasets were converted to secondary feature vector datasets of reduced dimension using the FastICA method. The secondary vector datasets were then fed to the QSL algorithm and the results were compared against the original experiments.

Two example ROC curves that were obtained from the original and the ICA experiments are given in Fig. 5. It can be observed that the ROC curves obtained from the ICA applied vector datasets correspond to better labelling performances compared to the original ROC curves. The labelling performance measures for the original and the ICA applied experiments are presented in Table 3. The results indicate that in all of the ICA operated experiments, except the one on the Dataset 6, the ICA method improved the labelling performances.

The sample outputs for the automated labelling of QSL method operated on a vector dataset of reduced dimensionality, overlaid

Table 3
Labelling performances of the QSL method for original and the ICA applied texture feature vector datasets.

Texture features	Original		ICA	
	(P_{FP}, P_{TP})	ROC area	(P_{FP}, P_{TP})	ROC area
Dataset 1	(0.25, 0.79)	0.84	(0.21, 0.84)	0.89
Dataset 2	(0.19, 0.84)	0.88	(0.22, 0.88)	0.90
Dataset 3	(0.19, 0.82)	0.88	(0.21, 0.86)	0.89
Dataset 4	(0.20, 0.83)	0.88	(0.18, 0.86)	0.91
Dataset 5	(0.21, 0.83)	0.87	(0.19, 0.88)	0.91
Dataset 6	(0.23, 0.84)	0.87	(0.32, 0.98)	0.86

on histopathology images are shown in Fig. 6(c) and (d). In these figures, individual square regions that were identically labelled were merged and framed as an overlay onto original histopathology images. The ground truth atlas data corresponding to these two histopathology images are presented in Fig. 6(a) and (b) can also be compared with these outputs.

4. Conclusions and discussion

In this paper, we have presented a quasi-supervised texture labelling methodology to recognize NNCR and CRCa textures in H&E stained colorectal histopathology images. The main contribution of the proposed method was significant reduction of the expert supervision before the automated recognition procedure. The assignment of the images in the Histopathological Image Library into NNCR and mixed collections was minimal compared to preparing the ground-truth atlas data using the HIAE.

In the experiments performed on colorectal cross-section images, the regions containing CRCa tissue were identified with high accuracy without using manually labelled ground-truth information. Since there is no histopathology image texture benchmark database providing a comparison of accuracies achieved by other classification schemes, it was unfortunately not possible to check our results with the other methods from the literature. Experimental results indicated that the probability of false positive values by the SVM classifier were relatively lower compared to those by the QSL algorithm. On the other hand, the probability of true positive values was very close for both methods.

It was also observed that among the various datasets corresponding to different feature extraction parameters, the output labelling performance of the QSL method was minimum when operated on Dataset 1 constructed using grey level image information. The colour texture feature information derived from the *Lab* colour space offered the best characterization of the NNCR and CRCa features.

In this study, we have experimented with several dimensionality reduction procedures to determine if reducing the texture vector dimensionalities would lead better classification performances. Among these procedures, only the ICA method improved the labelling performance. One of the possible reasons for the improvement is that ICA could extract the valuable vector component information despite their low variances and also suppresses
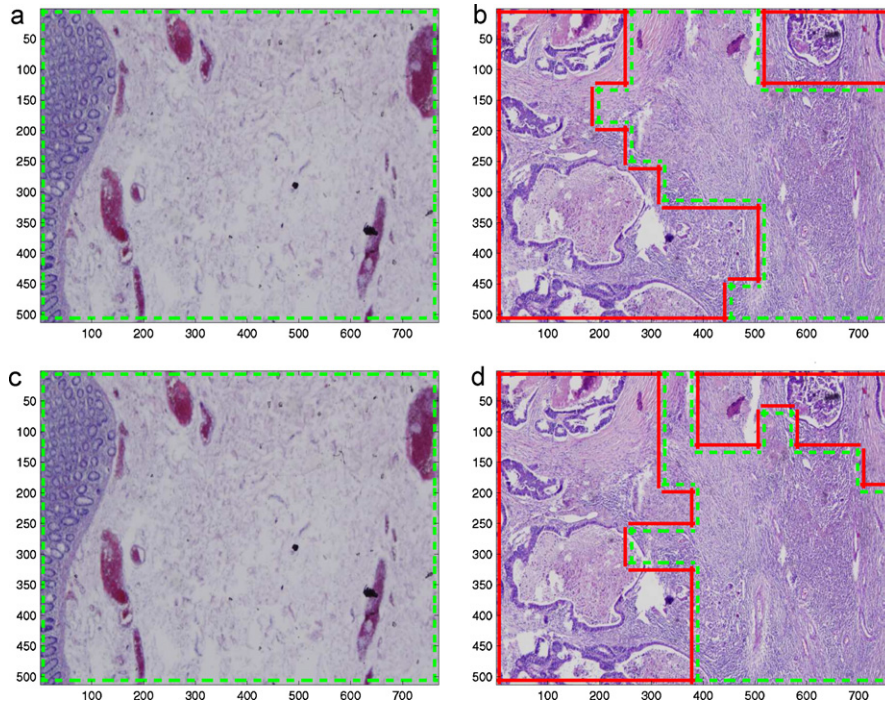
Fig. 6. QSL labelling results with the ground truth information. (a) and (b): Two histopathology images with overlaid ground truth atlas data. (a) Completely consists of NNCR and (b) has both NNCR and CRCa tissues. QSL labelling results following the ICA. (c) and (d): Regions bounded by dashed green lines imply NNCR and solid red lines imply CRCa tissue regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the redundant data. In majority of the ICA operated texture labelling experiments, the output recognition performances improved with the corresponding P_{TP} values approaching the upper bound obtained from the SVM classification experiments. On the other hand, the P_{FP} values still remained higher than those obtained by the SVM.

In this study, despite the high performance levels in terms of P_{TP}, we have faced a phenomenon referred to as "malign tendency" which came up with high false positive rates, implying a general bias towards labelling NNCR texture vectors as CRCa, observed especially in the mixed labelled test group. We believe that this phenomenon is due to severely complicated local structures in colonic tissues and the resulting weak separation of normal and cancer feature vectors in the multidimensional vector space.

Moreover, we have checked this assumption by performing several QSL experiments on synthetic vector data. We have generated statistically well differentiated random feature vectors for both NNCR and CRCa tissues in high dimensional space and obtained excellent QSL labelling performances. This evidence showed us that in order to increase the QSL accuracy, different texture feature characteristics must be incorporated into the automated labelling framework. A feature extraction scheme using co-occurrence matrices calculated by the pixel pairs of a specified orientation could be tested. There are also several alternatives to the features used in this study and the comparison of their labelling performances with the QSL accuracies remains to be evaluated for the optimum configuration.

The proposed framework and several MIL applications in histopathology should be specially compared. The main similarity is that the proposed framework can be converted into a MIL application by putting the vectors calculated from a reference group slide into a negative bag and putting the vectors calculated from a mixed group slide into a positive bag. This is the case in which positive bags do not contain any negatively labelled instances and a special MIL algorithm needs to be implemented to handle this. The main contrast between these two frameworks is that, in the MIL based

frameworks, the need for expert marking still exits during the classifier training phase (Zhao et al., 2006; Dundar et al., 2008, 2010, 2011).

Although there are MIL classifiers that calculate the output labels for each instance (Andrews et al., 2002; Chen et al., 2006) several MIL classifiers were designed to perform slide-level classification (Dundar et al., 2010, 2011). In contrast, the QSL algorithm calculates the posterior probabilities for all individual instances. It should also be pointed out that while many standard supervised learning methods, such as, K-Nearest-Neighbourhood, Neural Network, and Decision Trees, have been adopted to solve the MIL problem, the QSL method has been simply build upon a Nearest-Neighbourhood classifier which uses the Euclidian vector distances. Furthermore, the QSL framework does not have any constraint for bag or instance selection.

It should also be noted that no colour normalization method was applied to the Histopathological Image Library images due to fixed capture and illumination parameters during image acquisition. Thus, our assumption was that these images were assumed to share similar visual standards and lack artifacts that can occur due to variations in the sample preparation or image acquisition procedures. However, whether these images possessed a variation imperceptible to the naked eye in the H&E staining process that might be affecting the computerized analysis remains an open question. Another research can thus be carried out by performing a colour normalization method in a preprocessing step of the proposed framework prior to statistical analysis using the QSL method (Macenko et al., 2009; Magee et al., 2009; Ruifrok and Johnston, 2001; Ruifrok et al., 2003).

In this study, many dimensionality reduction procedures were experimented and the ICA method was found to be the only one improving the labelling accuracy. However, several other feature selection or extraction algorithms leading to dimensionality reduction could not be evaluated due to long computation times required by these methods for the texture feature vector datasets used in this study. In the future, these algorithms, such as Exhaustive Selection

(Jain and Zongker, 1997) and Branch and Bound method (Narendra and Fukunaga, 1977) can be applied to the existing feature vector datasets provided that they can be carried out within limits of computational feasibility.

As another future study, several other texture feature types, such as, features obtained from the segmentation of the histopathology images into chromatin-rich, stromal and unstained regions, wavelet transforms and fractal dimensions will be considered for NNCR and CRCa texture separation (Karacali and Tozeren, 2007; Jafari-Khouzani and Soltanian-Zadeh, 2003; Li et al., 2007; Zhang and Ma, 2007; Shirazi et al., 2000). In addition, the proposed framework will also be applied to several other differential histopathology cases in colorectal or other tissue sections. Furthermore, the proposed method can form the core part of a semi-automated texture recognition framework in order to support pathologists by detecting the differential textures. To this end, the development may also focus on a first step evaluation of huge number of histopathology whole-slide tissue cross-section images.

Acknowledgements

References

Andrews, S., Tsochantaridis, I., Hofmann, T., 2002. Support vector machines for multiple-instance learning. In: Neural Information Processing Systems, pp. 561–568.

Burges, J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167.

Chen, Y., Bi, J., Wang, J.Z., 2006. Miles: multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12), 1931–1947.

CIE, 1986. CIE colorimetry: official recommendations of the international commission on illumination. Tech. Rep., CIE.

Comon, P., 1994. Independent component analysis, a new concept? Signal Processing 36 (3), 287–314 (Special issue on Higher Order Statistics).

Cortes, C., Vapnik, V.N., 1995. Support-vector networks. Machine Learning 273–297.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. Information Theory 13, 21–27.

Diamond, D.A., Berry, S.J., Umbricht, C., Jewett, H.J., Coffey, D.S., 1982. Computerised image analysis of nuclear shape as a prognostic factor for prostatic cancer. Prostate 3, 321–332.

Dietterich, T.G., Lathrop, R.H., Lozano-Prez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89 (12), 31–71.

Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M., Tomaszeweski, J., 2007. Automated grading of prostate cancer using architectural and textural image features. In: 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2007), No. 9506433, Dept. of Biomed. Eng., State Univ. of New Jersey, New Brunswick, NJ, April 2007, pp. 1284–1287.

Doyle, S., Rodriguez, C., Madabhushi, A., Tomaszeweski, J., Feldman, M., 2006. Detecting prostatic adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach. In: 28th IEEE Annual International Conference of the Engineering in Medicine and Biology Society (EMBS'06), pp. 4759–4762.

Dundar, M., Badve, S., Bilgin, G., Raykar, V., Jain, R., Sertel, O., Gurcan, M., 2011. Computerized classification of intraductal breast lesions using histopathological images. IEEE Transactions on Biomedical Engineering 58 (July (7)), 1977–1984.

Dundar, M., Badve, S., Raykar, V., Jain, R., Sertel, O., Gurcan, M., 2010. A multiple instance learning approach toward optimal classification of pathology slides. In: 20th International Conference on Pattern Recognition (ICPR), August 2010, pp. 2732–2735.

Dundar, M., Fung, G., Krishnapuram, B., Rao, R., 2008. Multiple-instance learning algorithms for computer-aided detection. IEEE Transactions on Biomedical Engineering 55 (March (3)), 1015–1021.

Ficsor, L., Varga, V., Tagscherer, A., Tulassay, Z., Molnar, B., 2008. Automated classification of inflammation in colon histological sections based on digital microscopy and advanced image analysis. Cytometry 73 (Mar (3)), 230–237.

Filippas, J., Amin, S., Naguib, R., Bennett, M., 2003. A parallel implementation of a genetic algorithm for colonic tissue image classification. In: 4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, pp. 330–333.

Gonzalez, R., Woods, R., 1992. Digital Image Processing. Addison-Wesley, Reading, MA.

Hallouche, F., Adams, A.E., Hinton, O.R., Relf, G., Lakshmi, M.S., Sherbet, G.V., 1992. Image processing for cell cycle analysis and discrimination in metastatic variant cell lines of B16 murine melanoma. Pathobiology 60, 76–81.

Hamilton, P.W., Allen, D.C., Watt, P.C.H., 1987. Classification of normal colorectal mucosa and adenocarcinoma by morphometry. Histopathology 11, 901–911.

Hamilton, P.W., Bartels, P., Thompson, D., Anderson, N.H., Montironi, R., Sloan, J., 1997. Automated location of dysplastic fields in colorectal histology using image texture analysis. Journal of Pathology 182 (May (1)), 68–75.

Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3 (6), 610–621.

Huang, P.-W., Lee, C.-H., 2009. Automatic classification for pathological prostate images based on fractal analysis. IEEE Transactions on Medical Imaging 28 (7), 1037–1050.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis.

Jafari-Khouzani, K., Soltanian-Zadeh, H., 2003. Multiwavelet grading of pathological images of prostate. IEEE Transactions on Biomedical Engineering 50 (June (6)), 697–704.

Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 153–158.

Jolliffe, I., 2002. Principal Component Analysis.

Karacali, B., 2010. Quasi-supervised learning for biomedical data analysis. Pattern Recognition 43 (10), 3674–3682.

Karacali, B., Tozeren, A., 2007. Automated detection of regions of interest for tissue microarray experiments: an image texture analysis. BMC Medical Imaging 7 (1).

Kokturk, B.E., 2011. Separation of stimulus-specific patterns in electroencephalography data using quasi-supervised learning. Master's thesis. İzmir Institute of Technology.

Li, J., Mohamed, S.S., Salama, M.M.A., Freeman, G.H., 2007. Prostate tissue texture feature extraction for cancer recognition in TRUS images using wavelet decomposition. In: Kamel, M.S., Campilho, A.C. (Eds.), Proceedings of 4th International Conference on Image Analysis and Recognition, ICIAR 2007. Montreal, Canada, August 22–24, vol. 4633 of Lecture Notes in Computer Science. Springer, pp. 993–1004.

Macenko, M., Niethammer, M., Marron, J., Borland, D., Woosley, J., Guan, X., Schmitt, C., Thomas, N., 2009. A method for normalizing histology slides for quantitative analysis. In: Proceedings of the International Symposium on Biomedical Imaging (ISBI), pp. 1107–1110.

Magee, D., Treanor, D., Crellin, D., Shires, M., Mohee, K., Quirke, P., 2009. Colour Normalisation in Digital Histopathology Images.

Masood, K., Rajpoot, N., Qureshi, H., 2006. Co-occurrence and morphological analysis for colon tissue biopsy classification. In: Proceedings 4th International Workshop on Frontiers of Information Technology (FIT06), Islamabad.

Mookiah, M., Shah, P., Chakraborty, C., Ray, A., 2011. Brownian motion curve-based textural classification and its application in cancer diagnosis. Analytical & Quantitative Cytology & Histology 33 (3), 158–168.

Muthu, R.K., Vikram, V., Rajendra, A., Pal, M., Paul, R.R., Min, L.C., Ray, A.K., Chatterjee, J., Chakraborty, C., 2012. Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm. Micron 43, 352–364.

Narendra, P.M., Fukunaga, K., Sept 1977. A branch and bound algorithm for feature subset selection. IEEE Transactions on Computers 26 (9), 917–922.

Nasser Esgiar, A., Naguib, R.N.G., Sharif, B.S., Bennett, M.K., Murray, A., 1998. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. IEEE Transactions on Information Technology in Biomedicine 2 (3), 197–203.

Nwoye, E., Khor, L.C., Woo, W.L., Dlay, S.S., 2006. Spectral and statistical features in fuzzy neural expert machine for colorectal adenomas and adenocarcinoma classification. In: 5th International Symposium on Communication Systems, Networks and Digital Signal Processing, pp. 792–796.

Onder, D., Sarioglu, S., Karacali, B., 2010. Automated classification of cancerous textures in histology images using quasi-supervised learning algorithm. In: Biomedical Engineering Meeting, BIYOMUT 2010.

Pitts, D.E., Premkumar, Saganti, B., Houston, A.G., Babaian, R.J., Troncoso, P., 1993. Texture analysis of digitized prostate pathologic cross section. In: Proc. SPIE: Med. Imaging: Image Processing, SPIE, vol. 1898, pp. 465–470.

Pratt, R., 1991. Digital Image Processing. John Wiley, New York.

Rajpoot, K., Rajpoot, N., 2004. Hyperspectral colon tissue cell classification. In: SPIE Medical Imaging (MI).

Ruifrok, A.C., Johnston, D.A., 2001. Quantification of histochemical staining by color deconvolution. Analytical & Quantitative Cytology & Histology 23 (4), 291–299.

Ruifrok, A.C., Katz, R.L., Johnston, D.A., 2003. Comparison of quantification of histochemical staining by hue-saturation-intensity (HSI) transformation and color-deconvolution. Applied Immunohistochemistry & Molecular Morphology 11 (1), 85–91.

Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M., Adler, D., Goodsitt, M., 1996. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Transactions on Medical Imaging 15 (5), 598–610.

Sarioglu, S., Cakalagaoglu, F., Elagoz, S., Han, U., Etit, D., Hucumenoglu, S., Karaman, I., Koybasioglu, F., Kulacoglu, S., Paker, I.O., Ozbilim, G., Ozluk, Y., Pabuccuoglu, U., Ruacan, A., Seckin, S., Uguz, A., Saraydaroglu, O., Veral, A., Yilmazbayhan, D., Elli-dokuz, H., 2010. Inter-observer agreement in laryngeal pre-neoplastic lesions. Head and Neck Pathology 4, 281.

Schwarz, M., Cowan, W., Beatty, J., 1987. An experimental comparison of RGB, YIQ, LAB, HSV, and opponent color models. ACM Transactions on Graphics 6, 123–158.

Shirazi, M.N., Noda, H., Takao, N., 2000. Texture classification based on Markov modeling in wavelet feature space. Image and Vision Computing 18 (12), 967–973.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.

Waheed, S., Moffitt, R.A., Chaudry, Q., Young, A.N., Wang, M.D., 2007. Computer aided histopathological classification of cancer subtypes. In: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2007, October 2007, pp. 503–508.

Zhang, G., Ma, Z., 2007. Texture feature extraction and description using Gabor wavelet in content based medical image retrieval. In: International Conference on Wavelet Analysis and Pattern Recognition, vols. 1–4, pp. 169–173.

Zhao, D., Chen, Y., Correa, H., 2006. Automated classification of human histological images, a multiple-instance learning approach. In: Life Science Systems and Applications Workshop, IEEE/NLM, July 2006, pp. 1–2.