Expert Systems with Applications 39 (2012) 1062–1077

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Hybrid segmentation, characterization and classification of basal cell nuclei from histopathological images of normal oral mucosa and oral submucous fibrosis

M. Muthu Rama Krishnan [a], Chandan Chakraborty [a,*], Ranjan Rashmi Paul [b], Ajoy K. Ray [c]

[a] School of Medical Science and Technology, IIT Kharagpur, India
[b] Department of Oral and Maxillofacial Pathology, Guru Nanak Institute of Dental Science and Research, Kolkata, India
[c] Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, India

ARTICLE INFO

Keywords:
Oral submucous fibrosis
Zernike moments
Parabola fitting
Color deconvolution
Fuzzy divergence
Unsupervised feature selection
Gradient vector flow
Pattern classification

ABSTRACT

This work presents a quantitative microscopic approach for discriminating oral submucous fibrosis (OSF) from normal oral mucosa (NOM) in respect to morphological and textural properties of the basal cell nuclei. Practically, basal cells constitute the proliferative compartment (called basal layer) of the epithelium. In the context of histopathological evaluation, the morphometry and texture of basal nuclei are assumed to vary during malignant transformation according to onco-pathologists. In order to automate the pathological understanding, the basal layer is initially extracted from histopathological images of NOM ($n = 341$) and OSF ($n = 429$) samples using fuzzy divergence, morphological operations and parabola fitting followed by median filter-based noise reduction. Next, the nuclei are segmented from the layer using color deconvolution, marker-controlled watershed transform and gradient vector flow (GVF) active contour method. Eighteen morphological, 4 gray-level co-occurrence matrix (GLCM) based texture features and 1 intensity feature are quantized from five types of basal nuclei characteristics. Afterwards, unsupervised feature selection method is used to evaluate significant features and hence 18 are obtained as most discriminative out of 23. Finally, supervised and unsupervised classifiers are trained and tested with 18 features for the classification between normal and OSF samples. Experimental results are obtained and compared. It is observed that linear kernel based support vector machine (SVM) leads to 99.66% accuracy in comparison with Bayesian classifier (96.56%) and Gaussian mixture model (90.37%).

1. Introduction

Oral cancer (OC) is the sixth most common cancer in the world. It accounts for approximately 4% of all cancers and 2% of all cancer deaths world-wide. In India it is the commonest malignant neoplasm, accounting for 20–30% of all cancers (Banoczy, 1982; Burkhardt, 1985; Daftary et al., 1993). A higher incidence of OC is observed on the Indian subcontinent mainly due to the late diagnosis of potentially precancerous lesions. Oral submucous fibrosis (OSF) is an insidious chronic, progressive, precancerous condition with a high degree of malignant potentiality. A large number of these cases transform into OC. Through progression of this pathosis, OC develops in the epithelial region of the oral mucosa. The precancerous status is judged on the basis of light microscopic histopathological features of oral epithelial dysplasia (OED) and/or cellular atypia which have different grades according to involvement of the epithelial region (Paul et al., 2005).

There is no established quantitative technique by which histopathologically significant features of the diseased tissue like (i) thickness of different histological layers, (ii) density, distribution, and alignment of tissue components, and (iii) cell population density, distribution, and their different morphological attributes could be analyzed. Actually, a precancerous state generally depicts mixed features of normalcy as well as pro- or pre-malignancy. With the disease progression, the histological scenario alters slowly in different combinations, characterizing the specific pathological state of progression toward malignancy (Paul et al., 2005).

Pathologists have been using microscopic images to study tissue biopsies for a long time, relying on their personal experience on giving decisions about the healthiness state of the examined biopsy. This includes distinguishing normal from abnormal (i.e., cancerous) tissue, benign versus malignant tumors and identifying the level of tumor malignancy. Nevertheless, variability in the reported diagnosis may still occur (Duncan & Ayache, 2000), which could be due to, but not limited to the heterogeneous nature of the diseases; ambiguity caused by nuclei overlapping; noise arising from the staining process of the tissue samples; intra-observer variability, i.e., pathologists are not able to give the same reading of the same image at more than one occasion; and inter-observer

* Corresponding author. Address: School of Medical Science and Technology, Indian Institute of Technology Kharagpur, West Bengal 721 302, India. Tel.: +91 3222 283570; fax: +91 3222 28881.
E-mail address: chandanc@smst.iitkgp.ernet.in (C. Chakraborty).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.07.107

variability, i.e., increase in classification variation among pathologists. Therefore, over the past three decades, quantitative techniques have been developed for computer-aided diagnosis, which aims to avoid unnecessary biopsies and assist pathologists in the process of cancer diagnosis (Gilles et al., 2008; Grootscholten et al., 2008; Shuttleworth, Todman, Norrish, & Bennett, 2005). Thus, quantitative evaluation of the histopathological features is not only important for accurate diagnostics, but it is also vital for assessing the relative involvement of the different tissue components in the pathology of the disease.

Generally, basal cells form the proliferative compartment (Shabana, Gel-Labban, & Lee, 1987) of the epithelium from which cells migrate, differentiate as they progress and eventually desquamated at the surface. The keratinocytes of the basal cell layer of the oral stratified squamous epithelium represent the progenitor cells (Satheesh, Paul, & Hammond, 2007) that are responsible for the production of other cells making the various layers of the epithelium. Changes in the basal cells may have serious implications on future cell behavior, including malignant transformation. The measurement of their size and shape in OSF may be an important prognostic marker as studies have shown that there is an increase in the size and shape of both the cell and the nucleus during OSF.

Automatic grading of pathological images has been investigated in various fields during the past few years, including brain tumor as to cytomas (ASTs) (Glotsos, 2003; McKeown & Ramsey, 1996; Scarpelli, Bartels, Montironi, Galluzzi, & Thompson, 1994; Schad, Schmitt, Oberwittler, & Lorenz, 1987), prostate carcinoma (Farjam, Soltanian-Zadeh, Zorrofi, & Jafari-Khouzani, 2005; Jafari-Khouzani & Soltanian-Zadeh, 2003; Smith, Zajicek, Werman, Pizov, & Sherman, 1999; Tabesh et al., 2007), renal cell carcinoma (RCC) (Fuhrman, Lasky, & Limas, 1982; Hand & Broders, 1931; Kim, Choi, Cha, & Choi, 2005; Lohse, Blute, Zincke, Weaver, & Chenille, 2002; Novara, Martignoni, Artibani, & Ficarra, 2007), and hepatocellular carcinoma (Huang & Lai, 2010); however, an automated system for screening OSF biopsy images has not been exhaustively reported in the literature, but some works have been reported (Muthu Rama Krishnan et al., 2009; Muthu Rama Krishnan, Shah et al., 2010) in other layers of oral mucosa. Since the grading system for a specific type of cancer cannot be applied to other types of cancers, it is necessary to exploit appropriate segmentation, feature extraction, and classification methods for different types of cancers. This is particularly true for the oral cancer because OSF biopsy images always suffer from the problem of impurities, undesirable elements, and uneven exposure. In OSF screening, the characteristics of basal cell nuclei are the key to estimate the degree of oral malignancy. However, the areas of nuclei, cytoplasm, and cells are difficult to be identified and measured. In this paper, we propose a novel method to segment the basal cell nuclei. Twenty three features are extracted from OSF biopsy images according to five types of characteristics commonly adopted by pathologists.

We use set of supervised (Bayesian and SVM) and unsupervised (*k*-means, fuzzy *c*-means and Gaussian mixture model (GMM)) classifiers to test the effectiveness of classification for OSF biopsy images. In this study, we find that not all 23 features are equally important or necessary to distinguish normal and OSF images. Therefore, we implemented an unsupervised feature selection for optimal feature subset selection so that the best performance of classifying OSF images can be achieved.

2. Materials and methods

2.1. Histology

Twelve study subjects clinically diagnosed with OSF have been subjected to incisional biopsy under their informed consent at the Department of Oral and Maxillofacial Pathology, Guru Nanak Institute of Dental Sciences and Research, Kolkata, India. Normal study samples are collected from the buccal mucosa of 10 healthy volunteers without having any oral habits or any other known systemic diseases with prior written consent. The study subjects are of similar age (21–40 years) and food habits. This study is duly approved by the ethics review committee of the institute. All the biopsy samples processed for histopathological examination and paraffin embedded tissue sections of 5 μm thickness prepared and then stained by haematoxylin and eosin (H&E).

2.2. Image acquisition

Images of basal layer for Normal oral mucosa and OSF are optically grabbed by Zeiss Observer. Z1 Microscope from H&E stained histological sections under 100× objectives (N.A.1.4) at School of Medical Science & Technology. At a resolution of 0.24 μm and the pixel size of 0.06 μm. Image database for this analysis consist of 1194 cells, which are extracted from 341 normal and 429 OSF with dysplasia images. The grabbed images are digitized at 1388 × 1040 pixels and stored in a computer.

2.3. Image processing

The histopathological image of oral mucosa grabbed by Carl Zeiss microscope contains white and black pixels (noise) randomly. To remove the noise median filter is used (Gonzalez & Woods, 2002). The block diagram of the proposed methodology for quantitative evaluation of basal cell nuclei is as shown in Fig. 1.

2.4. Basal cell nuclei segmentation

The novel approach for basal cell nuclei analysis mainly consists of three stages. (1) Basal nuclei extraction, (2) feature extraction and (3) classification. Basal nuclei extraction is a four step mechanism, i.e. (a) extraction of lower boundary of epithelium, (b) parabola fitting to segment the basal layer (c) segmentation of basal cell using marker controlled watershed and (d) extraction of nuclei using gradient vector flow. Anatomically, basal layer is the first layer in epithelium and first mechanism of basal nuclei extraction keeps this promise by extracting epithelio-mesenchymal (EM) junction. The first step in this mechanism consists of morphological operations on H&E stained image to diminish the local maxima (cell boundaries and variation in collagen fibers) present in epithelium and connective tissue followed by enhancing the edges by anisotropic diffusion and generating binary image using fuzzy divergence (Chaira & Ray, 2003, 2009) based thresholding.

2.4.1. Edge enhancement using anisotropic diffusion

Anisotropic diffusion (Grieg, Kubler, Kikinis, & Jolesz, 1992) is a technique aiming at removing or smoothing of the homogeneous part of an image keeping the significant part of the image like edge, line and other details that are important for the image interpretation. The main idea of this approach is to embed the original image in a family of derived images obtained by convolving the original image with a Gaussian kernel having variance *t*, which is a scale-space parameter. Larger values of *t* correspond to coarser resolution and lower values correspond to fine resolution. This one parameter family of derived images can be represented as a solution of the heat conduction or diffusion equation. Mathematically, the anisotropic diffusion is defined as

$$\frac{\partial I}{\partial t} = div(c(x,y,t)\nabla I) = \nabla c.\nabla I + c(x,y,t)\Delta I, \tag{1}$$

where Δ denotes the Laplacian operator, ∇ denotes the gradient and $c(x,y,t)$ is the diffusion coefficient which controls the rate of

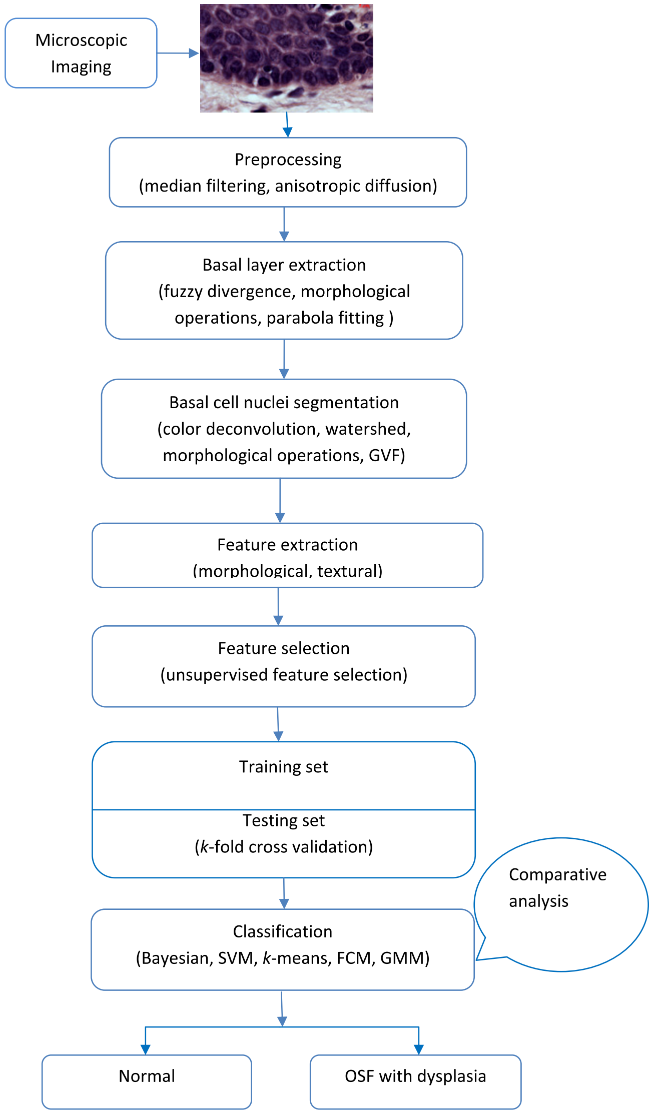
Microscopic Imaging
Preprocessing (median filtering, anisotropic diffusion)
Basal layer extraction (fuzzy divergence, morphological operations, parabola fitting)
Basal cell nuclei segmentation (color deconvolution, watershed, morphological operations, GVF)
Feature extraction (morphological, textural)
Feature selection (unsupervised feature selection)
Training set
Testing set (k-fold cross validation)
Comparative analysis
Classification (Bayesian, SVM, k-means, FCM, GMM)
Normal
OSF with dysplasia

Fig. 1. Block diagram of the proposed methodology.

diffusion. The solutions to the diffusion Eq. (1) is proposed by Perona and Malik (1990) as two functions for diffusion coefficients

$$c(\|\nabla I\|) = e^{\left(\frac{\|\nabla I\|}{K}\right)^2}, \tag{2}$$

$$c(\|\nabla I\|) = \frac{1}{1+\left(\frac{\|\nabla I\|}{K}\right)^2}, \tag{3}$$

the constant K controls the sensitivity to edges and the first privileges high-contrast edges over low-contrast edges and the second one privileges wide regions over smaller ones. 8-nearest neighbors' discretization of the Laplacian operator is used. After diffusion, the image edges are made sharp (Fig. 2(d)) it can be inferred from the histograms before and after diffusion. Further, thresholding is done using fuzzy divergence to segment out the surface epithelium.

2.4.2. Fuzzy divergence based threshold selection

Fan and Xie (1999) proposed fuzzy divergence from fuzzy exponential entropy by using a single row vector. Here the divergence concept of Fan and Xie is extended to an image, represented by a matrix. In an image of size $M \times M$ with L distinct gray level having probabilities $(p_0, p_1, p_2, \ldots, p_{L-1})$, the exponential entropy is defined as $H = \sum_{i=0}^{L-1} p_i e^{1-p_i}$.

The fuzzy entropy for an image A of size $M \times M$ is defined as

$$H(A) = \frac{1}{n(\sqrt{e}-1)} \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} \left[(\mu_A f_{ij}) \cdot e^{1-\mu_A f_{ij}} + (1-\mu_A f_{ij}) \cdot e^{\mu_A f_{ij}} - 1\right] \tag{4}$$

Here $n = M^2$ and $i, j = 0, 1, 2, 3, \ldots, (M-1)$. $\mu_A f_{ij}$ is the membership value if the pixel in the image and f_{ij} is the (i,j)th pixel of the image A. For two images A and B, at the (i,j)th pixel of the image, the information of discrimination between $\mu_A(a_{ij})$ and $\mu_B(b_{ij})$ of images A and B is given by (Chaira & Ray, 2003, 2009)

$$e^{\mu_A(a_{ij})}/e^{\mu_B(b_{ij})} = e^{\mu_A(a_{ij})-\mu_B(b_{ij})}, \tag{5}$$

where $\mu_A(a_{ij})$ and $\mu_B(b_{ij})$ are the membership values of the (i,j)th pixels in images A and B, respectively. $i, j = 0, 1, 2, \ldots, M-1$. The discrimination of image A against image B may be given as

$$D_1(A,B) = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} \left(1-(1-\mu_A(a_{ij}))e^{\mu_A(a_{ij})-\mu_B(b_{ij})} - \mu_A(a_{ij})e^{\mu_B(b_{ij})-\mu_A(a_{ij})}\right) \tag{6}$$

Likewise the discrimination of B against A is

$$D_2(B,A) = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} \left(1-(1-\mu_B(b_{ij}))e^{\mu_B(b_{ij})-\mu_A(a_{ij})} - \mu_B(b_{ij})e^{\mu_A(a_{ij})-\mu_B(b_{ij})}\right). \tag{7}$$

So, total fuzzy divergence between image A and B is obtained from Eqs. (6) and (7)

$$D(A,B) = D_1(A,B) + D_2(B,A), \tag{8}$$

$$D(A,B) = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} \Big(2 - (1-\mu_A(a_{ij}) + \mu_B(b_{ij})).e^{\mu_A(a_{ij})-\mu_B(b_{ij})} - (1-\mu_B(b_{ij}) + \mu_A(a_{ij})).e^{\mu_B(b_{ij})-\mu_A(a_{ij})}\Big). \tag{9}$$

In the method, image A is an original image and image B is an ideally segmented image. An ideally segmented image is defined as the image which is perfectly thresholded so that each pixel belongs to exactly either to the object or to the background region. In such situation, the membership values for ideally segmented image of each pixel belong to the object/background region should be equal to one. Hence the above Eq. (9) becomes,

$$D(A,B) = \sum_{i=o}^{M-1}\sum_{j=0}^{N-1} \left(2-(2-\mu_A(a_{ij})).e^{\mu_A(a_{ij})-1} - \mu_A(a_{ij}).e^{1-\mu_A(a_{ij})}\right). \tag{10}$$

Henceforth, in that way the divergence value of each pixel is calculated for whole image and corresponding gray level is noted. The gray value corresponding to the minimum divergence (Fig. 3(b)) is chosen as threshold initially for segmenting the object (epithelium) and background (rest of the image) regions. In fact, the minimum divergence value indicates the maximum belongingness of each object pixel to the object region (epithelium) and each background pixel to the background region (connective tissue). Morphological operations are performed on Fig. 3(c) to diminish the local maxima (cell boundaries and variation in collagen fibers) present in epithelium and connective tissue output is shown in Fig. 3(d). The area having maximum white pixels is extracted using connected component labeling (Fig. 3(e)).

Next, the edges, boundaries are extracted from this binary image using canny edge detector (Fig. 3 (f)). The longest edge present in this image is the epithelio-mesenchymal (EM) junction, which is

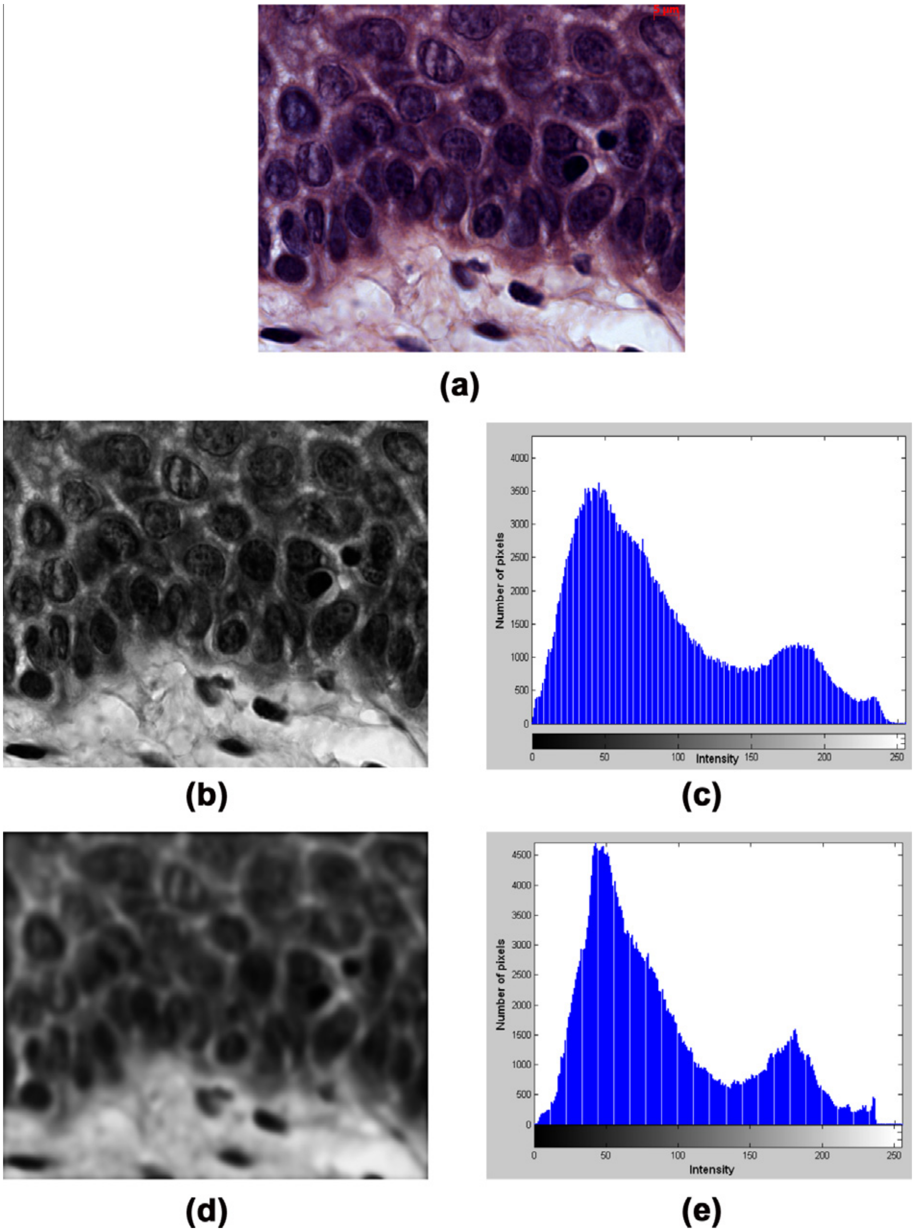

Fig. 2. (a) Normal colour image; (b) gray scale image of (a); (c) histogram of image (b); (d) diffused image of (b) using anisotropic diffusion; and (e) histogram of image (d).

extracted by 'Connected component labeling' to locate the lower boundary of the basal layer (Fig. 3 (f)). The abrupt variation in this edge is lessened by filtering it with band-pass filter. The shape and orientation of this lower boundary at 100× magnification can be approximated by parabola (Fig. 3(g)). Parabola fitting is performed by linear regression as parabola equation is a linear model (Rust, 2001).

2.4.3. Parabola fitting

Generalized equation for parabola is $Y = aX^2 + bX + C$. If the straight line model is inadequate for given data set, polynomial with degree 2, i.e., parabola may be one of the good choices as higher orders polynomial are unstable. Polynomial equation is a linear model so generalized model can be used to obtain linear regression (Muthu Rama Krishnan, Pal et al., 2010). The model for the $(n+1)$th order or nth degree polynomial is

$$y(t) = \sum_{i=1}^{n+1} \alpha_i X^{i-1}. \tag{11}$$

In matrix form,

$$\begin{bmatrix} y(t1) \\ y(t2) \\ y(t3) \\ \vdots \\ y(tm) \end{bmatrix} = \begin{bmatrix} 1 & X_{t1} & X_{t1}^2 & & X_{t1}^n \\ 1 & X_{t2} & X_{t2}^2 & \dots & X_{t2}^n \\ & \vdots & & \ddots & \vdots \\ 1 & X_{tm} & X_{tm}^2 & \dots & X_{tm}^n \end{bmatrix} \begin{bmatrix} \alpha \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_{n+1} \end{bmatrix}$$

or in shorter form, $Y = X\alpha$, where X is $m \times n+1$-dimensional matrix $(n \Leftarrow m)$ and α is a $n+1$ column vector.

Let us write the objective function for the least square estimation as

$$L(\alpha) = \sum_{i=1}^{m} [Y - X\alpha]_i^2 = [Y - X\alpha]^T [Y - X\alpha]. \tag{12}$$

Which we can expand to give

$$L(\alpha) = Y^T Y - 2\alpha^T X^T Y + \alpha^T X^T X\alpha. \tag{13}$$

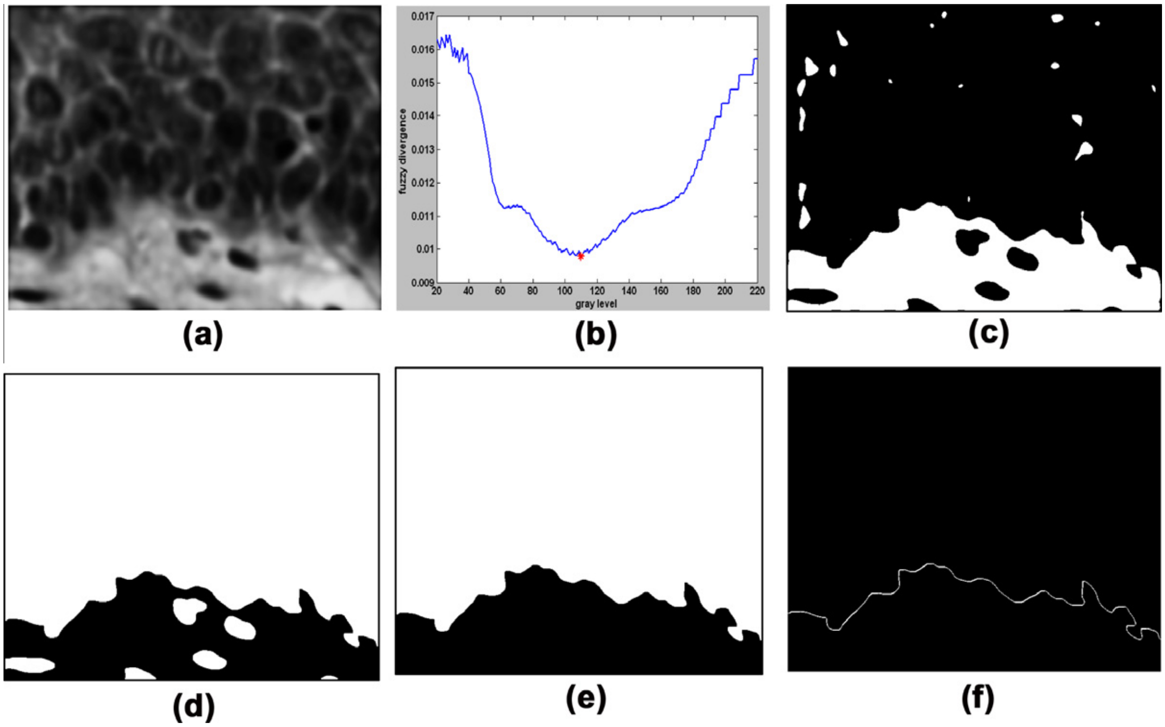

Fig. 3. (a) Normal gray scale image; (b) plots of gray level against fuzzy divergence, for selection of the threshold value; (c) thresholded image of (a); (d) morphological operation to remove small objects within the epithelium; (e) larger white area extracted using connected component labeling; and (f) lower boundary extraction from (e).

Geometrically, the objective function defines an $(n+1)$ dimensional, quadratic hypersurface, sometimes called the *response surface*, whose level curves correspond to concentric n-dimensional ellipsoids in the α-space. It has a unique global minimum that we can find by differentiating $L(\alpha)$ with respect to α and equating the result to the zero vector,

$$\frac{\partial L}{\partial \alpha} = -2X^TY + 2X^TX\alpha = 0.$$

Thus, the minimizing α must satisfy the $n \times n$ system of linear equations

$$X^TX\alpha = X^TY. \tag{14}$$

Which are often called the normal equations. Because the columns of X are linearly independent, the matrix product on the left side is nonsingular, so the unique solution is

$$\alpha = [X^TX]^{-1}X^TY. \tag{15}$$

If relatively small perturbations in the data produce relatively large perturbation in solution, we can get more numerically stable algorithm by computing an orthogonal factorization form

$$X = Q\begin{bmatrix} R \\ 0 \end{bmatrix},$$

where Q is an $m \times m$ orthogonal matrix $Q^TQ = I = QQ^T$, R is an $n \times n$ upper triangular matrix, and 0 is an $(m-n) \times n$ matrix of zeroes. By substituting this factorization into Equation, we can easily verify that α satisfies the $n\times$ n upper triangular system

$$R\alpha = Q_1Y, \tag{16}$$

where Q_1 is the $m \times n$ matrix formed by the first n columns of Q (Rust, 2001).

Assuming the model fitted to the data is correct, the residuals approximate the random errors. It is defined as $r_i = Y_i - X_i\alpha$, *for* $i = t_1, t_2, \ldots, , t_m$.

Therefore, if the residuals appear to behave randomly, it suggests that the model fits the data well. The parabola is fitted over the lower boundary (EM junction).

Next step is to generate 'n' parabola parallel to the fitted parabola for lower boundary of basal layer such that the image generated from these parallel parabola overlays basal layer completely (Fig. 4a and b). The effective distance between two parallel parabolas at distal end and center is not same. This property of the parallel parabolas generates image mask (Fig. 4(b)) and this mask is superimposed on Fig. 3(a), which gives basal layer (Fig. 4(c)).

2.4.4. Basal cell segmentation using color deconvolution

Moreover, epithelial cell borders cannot be isolated accurately in H&E stain; it can be estimated statistically using space partition procedure. Initially, the Haematoxylin plane is extracted using color deconvolution (Ruifrok & Johnston, 2001), which has high contrast between nuclei and cytoplasm.

2.4.4.1. Color deconvolution. According to Lambert–Beer's law, the detected intensities of light transmitted through the specimen and the amount (A) of stain with absorption factor c is described by

$$I = I_oe^{-Ac} \tag{17}$$

with I_0 is the intensity of light entering the specimen, I is the intensity of light detected after passing the specimen. This suggests that the gray-values of each RGB channel depend on concentration of stain in a non-linear way. Hence it is difficult to separate out each stain by intensity. However, the optical density (OD) can be used to separate it out and it is defined as

$$OD = -\log_{10}\frac{I}{I_o} = Ac. \tag{18}$$

Hence OD is proportional to absorption factor c for given amount of stain. This helps us to separate the contribution of each stain from multi stained specimen. Each pure stain will be characterized by a

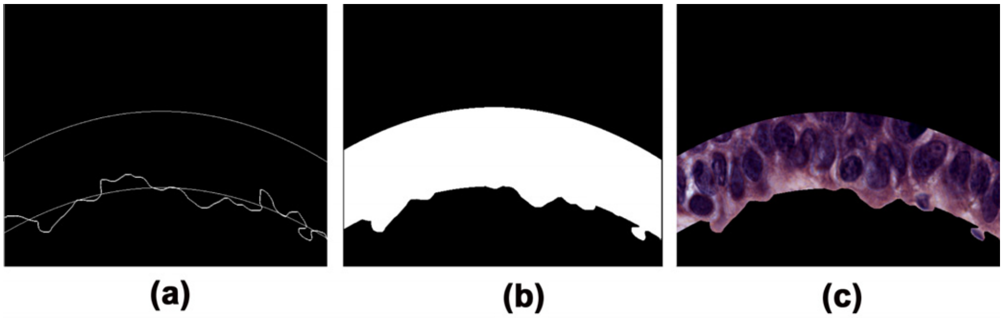

Fig. 4. (a) Parabola fitting using lower boundary; (b) mask of the fitted parabola; and (c) extracted basal layer using the mask.

specific optical density for the light in each of the three RGB channels, which can be represented by a 3 × 1 OD vector describing the stain in the OD-converted RGB color space. The length of the vector will be proportional to the amount of stain, while the relative values of the vector describe the actual OD for the detection channels (Ruifrok & Johnston, 2001).

In the case of three channels, the color system can be described as a matrix of the form with every row representing a specific stain, and every column representing the optical density as detected by the red, green and blue channel for each stain. Stain-specific values for the OD in each of the three channels can be easily determined by measuring relative absorption for red, green and blue on slides stained with a single stain.

If we can find out the ortho-normal transformation of this matrix, it is easy to separate each stain contribution. The transformation has to be normalized to achieve correct balancing of the absorbtion factor for each separate stain. If matrix M is normalized matrix of matrix OD then it is defined as

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix},$$

where $m_{ij} = O_{ij} \Big/ \sqrt{\sum_{k=1}^{3} O_{ik}^2}$ and O_{ij} is the element of the OD matrix.

If C is the 3 × 1 vector for amounts of the three stains at a particular pixel, then the vector of OD levels detected at that pixel is $y = MC$. From the above it is clear that $C = M^{-1}\,y$. This means, that multiplication of the OD image with the inverse of the OD matrix, which we define as the color-deconvolution matrix $D = M^{-1}$, results in orthonormal representation of the stains forming the image;

$$C = Dy. \tag{19}$$

This enhanced nuclei (Fig. 5(b)) with morphological operations (Fig. 5(d)) works as a marker in watershed algorithm to segment epithelium in different compartment. This compartment effectively shows the segmentation of epithelium in to 'basal cells' (Fig. 5(f)). Here, all partitions do not exactly contain the basal cells as some

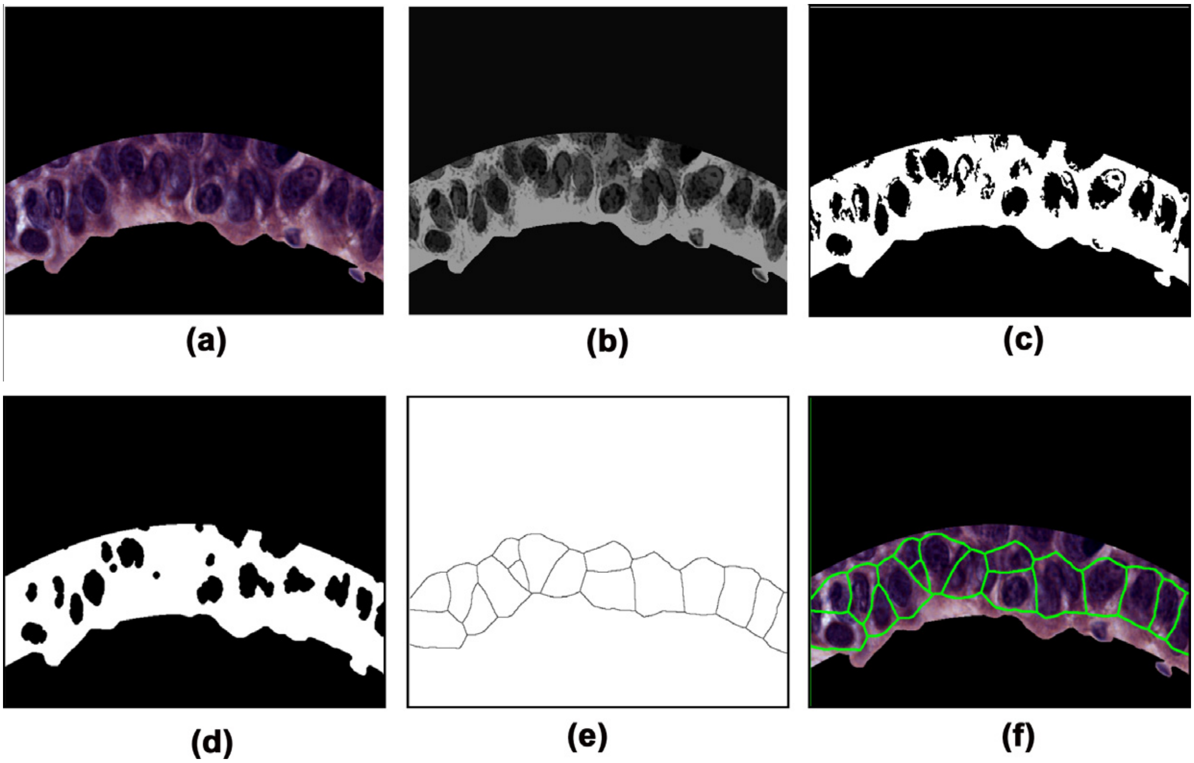

Fig. 5. (a) Extracted basal layer; (b) contrast enhanced nuclei using color deconvolution; (c) thresholded image of (b) using fuzzy divergence; (d) after performing morphological operations on image (c); (e) watershed output over image (d); (f) segmented boundaries of basal cells are superimposed on the extracted basal layer.

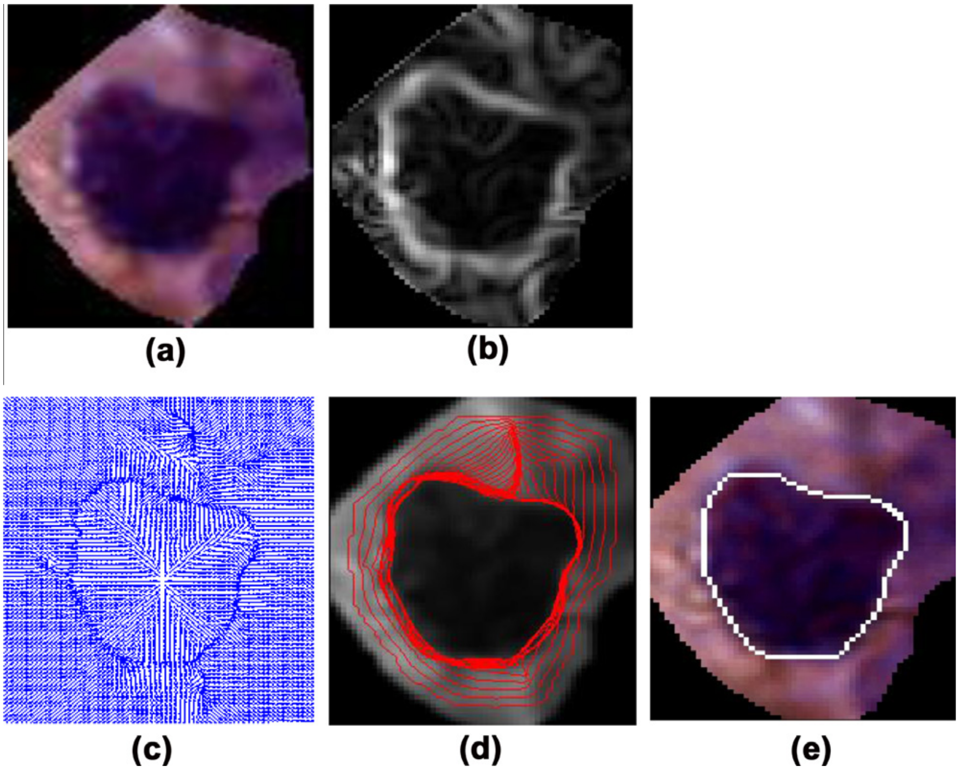

Fig. 6. (a) Basal cell image; (b) gradient image; (c) normalized GVF field; (d) deformation of the contour; and (e) final contour obtained using GVF.

of them have the suprabasal cells or clump of basal cells. The following approach is adopted to classify the partition so called pseudo cell as a basal cell or non-basal cell.

First step is to find the neighbors for all pseudo cells followed by evaluation of each pseudo cell area and if it is not within threshold, then it should be merged or ignored depending upon whether it is part of the cell or background respectively and named as ‘to be merged cell’. Further, shape parameter compactness and variance are evaluated for ‘to be merged cell’ and respective neighbor. These features are fuzzy in nature and are evaluated by trapezoid

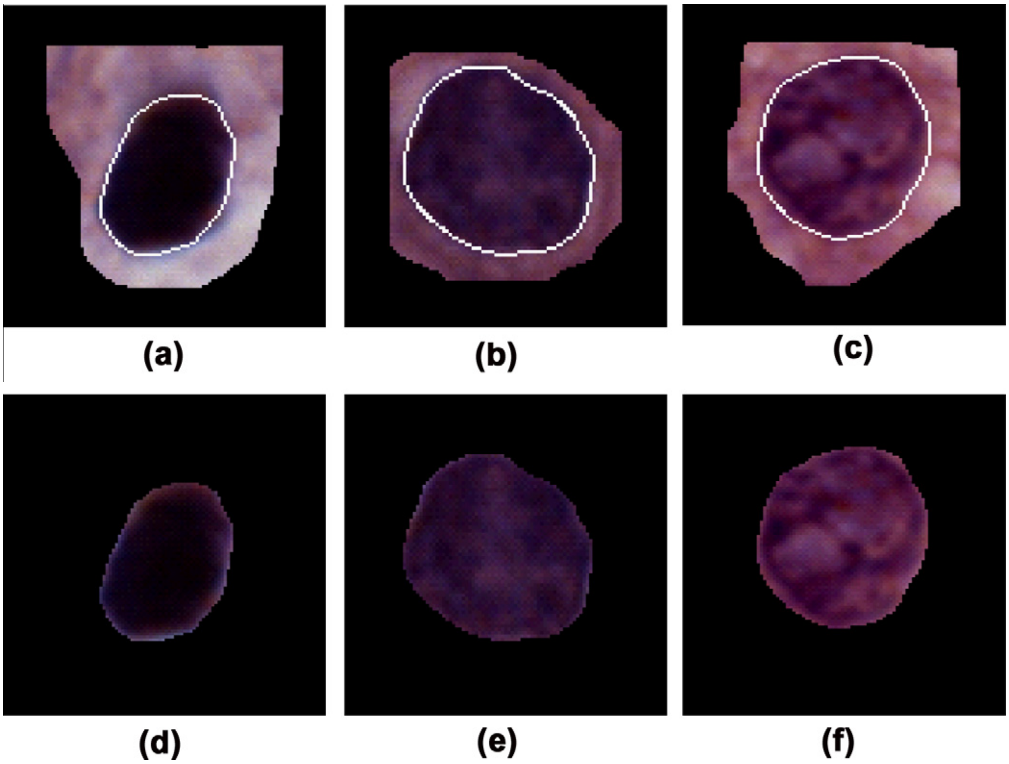

Fig. 7. (a–c) Basal cell nuclei contours tracked using GVF based snakes. (d–f) Segmented nucleus.

membership function. Then, 'to be merged cell' is merged with highest membership value. Moreover, the elimination of suprabasal layer is carried out by extracting the lowest cell from the image as basal layer is the first layer in epithelium.

2.5. Basal cell nuclei tracking using GVF snakes

The watershed segmentation gives the initial boundary around nuclei which also contains the background epithelial region. To segment the exact boundaries of objects, we use an energy -minimizing contour, called "snake" (Xu & Prince, 1997), which is guided by external constraint forces and influenced by image forces that pull towards the edges. Snake provides a powerful interactive tool for image segmentation. We use the contour obtained from the previous segmentation result as the initial contour, and then move this contour close to the more accurate nuclei contour under the influence of internal forces depending on the intrinsic properties of the curve and external forces derived from the image edge data.

To obtain good segmentation result, Gaussian Blur is applied on the image which restrains the noise in the cell image. The edge gradient of the image is computed using edge computation by Sobel operator. Flexible parameter α and rigid parameter β are also analysed by testing it. One different cases, which prove that the snake model cannot get good convergence result if α is less than 1. Hence α is taken to be 1.2. Furthermore, parameter β does not work in any cases. At the same time, iteration times of GVF Snake are analysed too. The segmentation result becomes stable if the iteration times for gradient computation is bigger than 80 in these cases.

Fig. 6(a–e) shows an example of a cell being tracked by the GVF snake. The red lines indicate the moving contour at different points of time. The segmented nuclei is shown in Fig. 7(d–f).

3. Feature extraction

The criteria of OSF with dysplasia grading are usually based on the following four types of characteristics: nuclear changes (variation in size and shape, polymorphism (nuclei of the basal layer are elongated and perpendicular to basement membrane), nuclear irregularity, hyperchromasia (excessive pigmentation in hemoglobin content of basal cell nuclei)). The above four types of characteristics are provided by experienced onco-pathologists (Paul et al., 2005) and usually used for OSF with dysplasia grading. In addition, to facilitate computer processing and image analysis, the onco-pathologists also suggest nuclear texture as the fifth type of characteristics. Then, 23 features based on these five types of characteristics are extracted from oral histopathological images for classification.

The following features are evaluated for nucleus. (a) Area, (b) perimeter, (c) eccentricity, (d) area equivalent diameter, (e) perimeter equivalent diameter, (f) convex area, (g) Zernike moments, and (h) Fourier descriptors, etc. Counting the number of pixels present in binary image of the nucleus gives the (f_1) *area*, whereas (f_2) *perimeter* of the nucleus has been obtained by counting the number of boundary pixels present in the nucleus. (f_3) *Form factor* is proportional to the area of each nucleus divided by the square of perimeter (http://www.dentistry.bham.ac.uk/landinig/software/software.html).

(f_3) *Form factor* mathematically defined as

$$\mathit{Form\ factor} = \frac{4 \times \pi \times \mathit{area}}{(\mathit{perimeter})^2}. \tag{20}$$

(f_4) Area equivalent diameter (AED) mathematically defined as

$$\text{Area equivalent diameter} = \sqrt{\frac{4}{\pi} \times \mathit{area}} \tag{21}$$

(f_5) Perimeter equivalent diameter (PED) mathematically defined as

$$\text{Perimeter equivalent diameter} = \sqrt{\frac{\mathit{area}}{\pi}}, \tag{22}$$

(f_6) *Eccentricity* is calculated by the following equation:

$$\text{Eccentricity} = \frac{\sqrt{a^2 - b^2}}{a}, \tag{23}$$

where a and b indicates major and minor axis. Which are obtained by elliptical approximation. Each nucleus has been approximated by a minimum bounding rectangle. The ellipse approximation has been done by first fitting the nucleus by a minimum bounding rectangle. The algorithm for minimum bounding rectangle is given in Chaudari and Samal (2007).

(f_7) Aspect ratio

It is mathematically defined as

$$\mathit{Aspect\ Ratio} = \frac{\mathit{Feret}}{\mathit{Breadth}}, \tag{24}$$

where *Feret*: Largest axis length of minimum bounding rectangle; *Breadth*: The largest axis perpendicular to the *Feret* (not necessarily colinear).

(f_8) Zernike moment.

The Zernike polynomials are first proposed in 1934 by Zernike. Their moment formulation appears to be one of the most popular, outperforming in terms of noise resilience, information redundancy and reconstruction capability. Complex Zernike moments are constructed using a set of complex polynomials which form a complete orthogonal basis set defined on the unit disc. They are expressed as Two dimensional Zernike moment (Khotanzad & Hong, 1990):

$$A_{mn} = \frac{m+1}{\pi} \int_x \int_y f(x,y)[V_{mn}(x,y)]^* \, dxdy, \tag{25}$$

where $m = 0, 1, 2,\ldots$ defines order of the moment and $f(x, y)$ is the function being described. Here n is an integer that depicting the angular dependence or rotation subject to the following condition:

$$m - |n| = even, \quad |n| \leqslant m. \tag{26}$$

Now, its expression in polar coordinates is

$$V_{mn}(r,\theta) = R_{mn}(r)\exp(jn\theta). \tag{27}$$

Here R_{mn} is the orthogonal radial polynomial and is defined as

$$R_{mn}(r) = \sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s F(m,n,s,r), \tag{28}$$

$$\text{and} \quad F(m,n,s,r) = \frac{(m-s)!}{s!\left(\frac{m-|n|}{2} - s\right)!\left(\frac{m-|n|}{2} + s\right)!} r^{m-2s}. \tag{29}$$

For a discrete case such as image, if $p(x, y)$ is the current pixel, the expression of Zernike moment becomes

$$A_{mn} = \frac{m+1}{\pi} \sum_x \sum_y p(x,y)[V_{mn}(x,y)]^*. \tag{30}$$

To calculate the Zernike moment, the image is first mapped to the unit disc using polar coordinates, where the centre of the image is the origin of the unit disc. Those pixels falling outside the unit disc are not used in the calculation. The coordinates are then described by the length of the vector from the origin to the coordinate point r and the angle from the x-axis to the vector r. Zernike moments are rotation invariants but not invariants to scaling and translation. Scaling and translation invariant can be achieved by transforming the pixel coordinate using following rule before applying Zernike moment

$$h(x,y) = f\left(\frac{x}{\alpha} + \bar{x}, \frac{y}{\alpha} + \bar{y}\right) \quad \text{where } \alpha = \sqrt{\frac{\beta}{m_{00}}} \tag{31}$$

and, $\bar{x} = \frac{m_{10}}{m_{00}}, \; \bar{y} = \frac{m_{01}}{m_{00}}$.
Here, m_{01}, m_{00}, m_{10} are the regular moments

$$m_{pq} = \sum_{x}\sum_{y} x^p y^q f(x,y). \tag{32}$$

Translation invariance is achieved by moving the origin to the image object center, causing $m_{01} = m_{10} = 0$. Following this, scale invariance is produced by altering each object so that its area (or pixel count for a binary image) is $m_{00} = \beta$, where β is a predetermined value (Khotanzad & Hong, 1990).

(f_9) Fourier descriptors

In any image (x_i, y_i) where $i = 1, 2, \ldots, K$ represents the edge points of an object, Fourier descriptors of that edge can be represented by the following approach. Each point can be treated as a complex number (Gonzalez & Woods, 2002) so that

$$s(k) = x_i + jy_i. \tag{33}$$

Now the DFT of $s(k)$ is

$$a(u) = \sum_{k=0}^{K-1} s(k)e^{-j2\pi uk/K}. \tag{34}$$

If we consider length of DFT of any sequence is same as original sequence, the total number of the descriptors varies as the length of the edge changes. Here the AC power of the Fourier descriptor is computed as follows:

$$P_{AC} = \sum_{u\neq 0, v\neq 0} (F_R^2(u,v) + F_I^2(u,v)), \tag{35}$$

where $F_R(u, v)$ and $F_I(u, v)$are real and imaginary parts of the Fourier transform of the image respectively, and u and v are the frequencies along the x and y axes of the image, respectively. Fourier descriptors are not invariants to scaling and translation. Scaling and translation invariant can be achieved using Eqs. (31) and (32).

(f_{10}) *Rectangularity* of the nuclei region mathematically defined as

$$R = \frac{A}{W \times H}. \tag{36}$$

A stands for the area, *W* stands for the width, *H* stands for the height. *R* describes the deviation degree of the nuclei to the rectangle.

(f_{11}) *Convex area*: Area of the convex hull (area of the smallest convex set of pixels containing the entire nuclear object).

(f_{12}) *Solidity*: Nuclear area divided by the area of the convex hull.

(f_{13}) *Roundness*: Nuclear area divided by the area of a circle with diameter equal to the length of the major axis.

(f_{14})

$$\textit{Concavity}: \textit{Convex Area-Nuclear Area}. \tag{37}$$

(f_{15}) *Orientation*: Angle (in degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.

(f_{16}) *Area Irregularity*: The nucleus is rotated so that its major axis becomes horizontal & is then enclosed by a minimum bounding rectangle (MBR). There is at least one intersecting point between a nucleus and each side of its MBR as shown in Fig. 8. If there are two or more intersecting points at one side, the middle one is selected as the representative intersecting point (Huang & Lai, 2010). Then, a nucleus is partitioned into four parts as follows. If an intersecting point is on a vertical side of the MBR, a horizontal cutting line will go through this point. If an intersecting point is on a horizontal side of the MBR, a vertical cutting line will go through this point. Consequently, four possibly overlapping areas S1, S2, S3, and S4 will be formed with each area surrounded by a segment of nucleus's boundary, a vertical line, and a horizontal line. The area irregularity is given as

$$\textit{Area Irregularity} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{4}\sum_{j=1}^{n}\max_{k=1\ldots4,k\neq j}\left|\left(\|S_j^i\| - \|S_k^i\|\right)\right|\right). \tag{38}$$

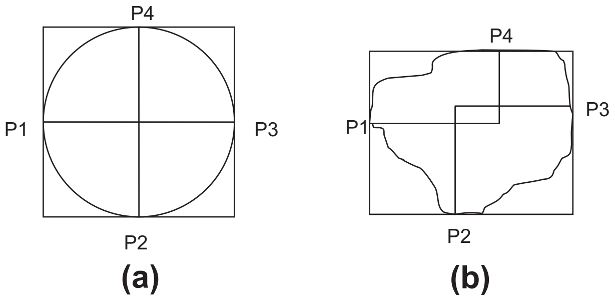

Fig. 8. Area irregularity (a) Round nucleus. (b) Irregular nucleus.

(f_{17}) *Contour irregularity*: The contour of the nucleus can be represented by a sequence of k equal spacing sample boundary points $\{p_0, \; p_1, \; p_2, \ldots, p_{j-1}p_j, \ldots, p_{k-1}\}$ with $p_k = p_0$ and $p_{-1} = p_{k-1}$. Let $p_j(w)$ be the boundary point with a distance of w pixels from the current point p_j. The curvature at point p_j is defined as:

$$d_j^i = \tan^{-1}\frac{y_j - y_j(w)}{x_j - x_j(w)} - \tan^{-1}\frac{y_{j-1} - y_{j-1}(w)}{x_{j-1} - x_{j-1}(w)}, \quad d_{-1}^i = d_{k-1}^i$$

Therefore, contour irregularity is defined as

$$\textit{Contour Irregularity} = \frac{1}{k}\left(\sum_{j=0}^{k-1}|d_j^i - d_{j-1}^i|\right), \tag{39}$$

(f_{18}) *Spot areas ratio*: Pigmentation is an important characteristic appearing in a malignant tumor. In our system, the bright and dark spots can be detected by top-hat and bottom-hat transforms, respectively, on nuclei using a disk shape structuring element of radius 5 (Huang & Lai, 2010). Top-hat transform is the difference between an input image and its opening by some structuring element, and bottom-hat transform is the difference between the closing and the input image. Top-hat transform returns an image containing elements that are smaller than the structuring element and brighter than their surroundings. Bottom-hat transform returns an image containing elements that are smaller than the structuring elements and darker than their surroundings.

$$\textit{Spot areas ratio} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\|N_i\|}(\|B(N_i)\| + \|D(N_i)\|)\right), \tag{40}$$

where $B(N_i)$: the overall size of all bright-spots innucleus N_i; $D(N_i)$: the overallsize of all dark-spots innucleus N_i; N_i: ith nucleus in the image $1 \leqslant i \leqslant n$.

Texture features: Haralick's texture features (Haralick, Shanmugan, & Dinstein, 1973) are calculated using the gray-level co-occurrence matrix. This matrix is square with dimension N_g, where N_g is the number of gray levels in the image. Element $[I, j]$ of the matrix is generated by counting the number of times a pixel with value I is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value I will be found adjacent to a pixel of value j. Four statistics namely (f_{19}) *contrast*, (f_{20}) *correlation*, (f_{21}) *homogeneity* and (f_{22}) *energy* are calculated from the co-occurrence matrices calculated using offsets as (1, 0); (−1, 0); (0, 1); (0, −1). Thus

$$Contrast = \sum_{i,j} |i-j|^2 p(i,j), \tag{41}$$

$$Correlation = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j}, \tag{42}$$

$$Homogeneity = \sum_{i,j} \frac{p(i,j)}{1+|i-j|}, \tag{43}$$

$$Energy = \sum_{i,j} p(i,j)^2. \tag{44}$$

(f_{23}) Hyperchromatism: Hyperchromatism represents excessive pigmentation in hemoglobin content of basal cell nuclei (Huang & Lai, 2010). It is an important characteristic appearing in a malignant tumor. For the case of sever dysplasia, chromatin abnormality will result in increasing staining capacity of nuclei. Thus, the intensity of nucleus in severe dysplasia usually appears darker than that of normal nucleus. To find the hyperchromatism mean intensity of nuclei (MNI) is calculated as follows:

$$\text{Mean intensity of nuclei} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\|N_i\|}\sum_{\forall (x,y)\in N_i} N_i(x,y)\right), \tag{45}$$

where n total number of nuclei, $N_{i:}$ ith nucleus in the image $1 \leqslant i \leqslant n$.

4. Unsupervised feature selection

All extracted features are checked for possibly highly correlated features. This process assists in removing any bias towards certain features which might afterwards affect the classification procedure. An approach which is based on feature similarity for measuring similarity between two random variables based on linear dependency (Mitra, Murthy, & Pal, 2002) proposed a measure called maximal information compression index. Let $\sum$ be the covariance matrix of random variable x and y. Define, maximal information compression index as $\lambda_2(x,y) = \textit{smallest eigenvalue of}\,\Sigma$, i.e.,

$$2\lambda_2(x,y) = \left(\mathrm{var}(x)+\mathrm{var}(y)-\sqrt{(\mathrm{var}(x)+\mathrm{var}(y))^2-4\mathrm{var}(x)\mathrm{var}(y)(1-\rho(x,y)^2)}\right). \tag{46}$$

The value of λ_2 is zero when the features are linearly dependent and increases as the amount of dependency decreases. It may not be noted that the measure λ_2 is nothing but the eigenvalue for the direction normal to the principal component direction of feature pair (x, y). It is shown that maximum information compression achieved if a multivariate data is projected along its principal component direction. The corresponding loss of information in reconstruction of the pattern (in terms of second order statistics) is equal to the eigenvalue along the direction normal to the principal component. Hence, λ_2 is the amount of reconstruction error committed if the data is projected to a reduced dimension in the best possible way. Therefore, it is a measure of the minimum amount of information loss or the maximum amount of information compression. The feature selection involves two steps, namely, partitioning the original feature set into a number of homogenous subsets (clusters) and selecting a representative feature from each such cluster. Partitioning of the features is done based on the k-NN principle using maximal information compression index. In doing so, we first compute the k nearest features of each feature. Among them the feature having the most compact subset (as determined by its distance to the farthest neighbor) is selected, and its k neighboring features are discarded. This process is repeated for the remaining features until all of them are either selected or discarded.

While determining the k nearest-neighbors of features, we assign a constant error threshold (ε) which is set equal to the distance of the kth nearest-neighbor of the feature selected in the first iteration. In subsequent iterations, we check the λ_2 value, corresponding to the subset of a feature, whether it is greater than ε or not. If yes, then we decrease the value of k.

5. k-Fold cross validation

k-Fold cross validation the data set is divided into k subsets. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set $k-1$ times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that we can independently choose how large each test set is and how many trials we average over (http://www.cs.cmu.edu/~schneide/tut5/node42.html).

6. Basal cell nuclei classification

The performance of our automatic basal cell nuclei classification system in this study is evaluated by two supervised and three unsupervised classifiers: the Bayesian classifier, the support vector machine (SVM) classifier, the k-means, the Fuzzy c-means and GMM clustering.

6.1. Bayesian classification

Bayesian classification is based on probability theory and the fundamental approach to the problem of classification is Bayes' decision theory (Duda, Hart, & Stork, 2007). The principle of the decision is to choose the most probable or the lowest risk (expected cost) option. The feature vector $x = [x_1, x_2, \ldots, x_d]$ is assumed to be generated by a d dimensional Gaussian process having ensemble mean μ and covariance matrix Σ such a process is represented using the probability density function given by

$$p(x_i|\lambda_k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\sum_k|^{\frac{1}{2}}} \exp\left\{\frac{-1}{2}(x_i-\mu_k)^T \sum^{-1}(x_i-\mu_k)\right\}. \tag{47}$$

The posterior probability of such process is computed by Bayes rule,

$$P(k|x_n) = \frac{\alpha_k p(x_n|\lambda_k)}{\sum_{j=1}^{c}\alpha_j p(x_n|\lambda_j)}, \tag{48}$$

where c is the number of classes present in the data and α_j is the jth class priori probability (>0). Here we have $c = 2$ viz., normal and OSF without dysplasia. In order to make a Bayesian decision, the following classification rule is adopted, If $P(k = 1|x_n) > P(k = 2|x_n)$ then $x_n \in Normal\ class$ else $x_n \in$ Osf class.

6.2. Support vector machine classification

The support vector machine classifier (El-Naqa, Yang, Wernick, Galatsanos, & Nishikawa, 2002; Vapnik, 1998) is based on the idea of margin maximization and it can be found by solving the following optimization problem

$$\begin{aligned} &\min \quad \frac{1}{2}w^T w + C\sum_{i=1}^{l} \xi_i^2 \\ &\text{s.t.} \quad y_i(w^T x_i + b) \geqslant 1 - \xi_i, i = 1, l, \xi_i \geqslant 0. \end{aligned} \tag{49}$$

The decision function for linear SVMs is given as $f(x) = w^T x + b$. In this formulation; we have the training data set $\{x_i, y_i\} i = 1, \ldots, l$, where $x_i \in R^n$ are the training data points or the tissue sample vectors, $y_i$are the class labels, l is the number of samples and n is the number of features in each sample. By solving the optimization problem (49), i.e., by finding the parameters w and b for a given training set, we are effectively designing a decision hyperplane over an n dimensional input space that produces the maximal margin in the space. Generally, the optimization problem (50) is solved by changing it into the dual problem below:

$$\max \qquad L_d(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j x_i^T x_j \tag{50}$$

Subject to $0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, l$

$$\sum_{i=1}^{l} \alpha_i y_i = 0. \tag{51}$$

In this setting, one needs to maximize the dual objective function $L_d(\alpha)$ with respect to the dual variables α_i only. Subject only to the box constraints $0 \leqslant \alpha_i \leqslant C$. The optimization problem can be solved by various established techniques for solving general quadratic programming problems with inequality constraints.

6.3. k-means clustering

The k-means clustering algorithm initially assumes k centroids (in our case $k = 2$). Based on the initial centroids, it calculates the cluster label to each pattern (consisting of 18 features) based on the minimum Euclidean distance (MacQueen, 1967). Based on these labels the centroids are re-estimated as the average of all the patterns belonging to that class at that iteration. The convergence criterion is total mean squared error that should be below a threshold. The iterations are continued until the total MSE is below the threshold. The k-means clustering minimizes following objective function.

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} \|x_i - c_k\|^2, \tag{52}$$

where x_i is the ith pattern and c_k is the kth centroid.

6.4. Fuzzy c-means clustering

The fuzzy c-means clustering algorithm optimizes (Bezdek, 1981) following objective function

$$J = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ji}^m \|x_i - V_j\|^2, \tag{53}$$

where u_{ji} is the fuzzy membership having m as the weighting exponent and with pattern x_i such that it can associate with the cluster j having centroid V_j. The fuzzy membership has the property such that

$$\sum_{j=1}^{c} u_{ji} = 1 \quad \forall i. \tag{54}$$

The algorithm almost works in the same manner as that of k means algorithm. The update equations for the Cluster center and the fuzzy membership are follows:

$$V_j^{(new)} = \frac{\sum_{i=1}^{N} (u_{ji})^m x_i}{\sum_{i=1}^{N} (u_{ji})^m} \tag{55}$$

$$u_{ji}^{(new)} = \frac{\left(\frac{1}{\|x_i - V_j^{(new)}\|}\right)^{\frac{2}{(m-1)}}}{\sum_{l=1}^{c} \left(\frac{1}{\|x_i - V_l^{(new)}\|}\right)^{\frac{2}{(m-1)}}} \tag{56}$$

The iterations are stopped when $\|U^{(new)} - U\|_F < \varepsilon$, a predefined small real number and $U = \{u_{ji}, 1 \leqslant j \leqslant c, 1 \leqslant i \leqslant N\}$.

For different weighting exponent, it is possible to get different clustering accuracies.

6.5. Gaussian mixture model based clustering

Here we have a binary class problem of classification of normal and OSF with dysplasia cases. The GMM (Bilmes, 1998) assumes that the features are drawn from a normal distribution. We have two mixing components corresponding to normal and OSF classes respectively. Therefore we have two class conditional densities, $p(x_n|\omega_k)$, $1 \leqslant k \leqslant 2$ and $1 \leqslant n \leqslant N$, where k is the number of classes and N is the total number of observations or patterns, corresponding class prior probabilities, $p(\omega_k)$, $1 \leqslant k \leqslant 2$. Each of the two mixing component has a mean vector and covariance matrix. Since we have applied orthogonal transformation in compact supported basis, the off diagonal elements in the covariance matrix are all approximately zero since the data will be highly uncorrelated. The probability density function of such a model is given by

$$p(x_n|\omega_k) = \frac{1}{2\pi|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x_n - \bar{x}_k)^T \Sigma_k^{-1}(x_n - \bar{x}_k)\right\}, \tag{57}$$

$$\text{where} \quad \bar{x}_k = \frac{1}{|X_k|} \sum_{x_n \in \omega_k} x_n, \tag{58}$$

$$\text{and} \quad \Sigma_k = \frac{1}{|X_k|} \sum_{x_n \in \omega_k} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^T = diag(\sigma_i^2), 1 \leqslant i \leqslant d. \tag{59}$$

The corresponding posterior probabilities are given by Bayes' rule as follows.

$$P(\omega_k|x_i) = \frac{p(x_i|\omega_k)}{\sum_{k=1}^{2} p(\omega_k) p(x_i|\omega_k)}. \tag{60}$$

Since our data consists of missing observations or it does not represent the whole of the sample space, the mean vectors and the covariance matrices computed are not the correct ones. Therefore the means and variances are recomputed using Expectation Maximization (EM) algorithm and using maximum likelihood estimation method. The re-estimating formulae are following

$$\hat{\mu}_j = \frac{\sum_{i=1}^{N} x_i P(\omega_j|x_i)}{\sum_{i=1}^{N} P(\omega_j|x_i)}, \tag{61}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu}_j)^2 P(\omega_j|x_i)}{\sum_{i=1}^{N} P(\omega_j|x_i)}, \tag{62}$$

$$p(\hat{\omega}_j) = \frac{1}{N}\sum_{i=1}^{N} P(\omega_j|x_i). \tag{63}$$

The initial prior probability is taken to be 0.5 for each of the classes. An initial model is assumed from the data. The EM algorithm used is having two core steps; E step and M step. During E step class conditional density is computed according to Eq. (57), and from it posterior density according to Eq. (60) is computed. During M step the class model is been re-estimated according to the Eqs. (61)–(63). The process is continued until the new estimate will not change much from the previous estimate, and model gets stabilized. Then the EM based GMM is said to be converged. The logarithm of the class conditional density called as log- likelihood is computed for each of the iteration and it will stop increasing at convergence.

The GMM algorithm is an optimization problem which maximizes the following objective function.

$$J = \prod_n \sum_k p(\omega_k)p(x_n|\omega_k) \tag{64}$$

The converged centroids are such that the product over all the observations, the total class conditional densities weighted with respective prior probability will be maximized. The EM algorithm determines its new estimate such that it will be approaching to the optimum of the objective function, so as for the algorithm to converge. GMM is an iterative algorithm, which can be performed in $O(ndkT)$ floating point operations, where n is the number of patterns, d is the total number of features in a pattern, k is the total number of classes present in the data, and T is the number of iterations required for convergence of the algorithm.

6.6. Performance analysis

In practice, each of the classifiers is required to be evaluated in order to compare their sensitivity, specificity along with overall accuracy. In view of this, the following confusion matrix (see Table 1) is usually designed based on the trade-off between actual and classifier generated outputs.

where

TP: True Positive: A patient predicted with OSF when the subject actually has OSF.

TN: True Negative: A patient predicted healthy when subject actually is healthy.

FP: False Positive: A patient predicted with OSF when subject actually is healthy.

FN: False Negative: A patient predicted healthy when subject actually has OSF.

Sensitivity: It is a measure of accuracy of diagnosis of malignant (true) cases of OSF. Mathematically, it is defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN}\%. \tag{65}$$

Specificity: It is a measure of accuracy of diagnosis of benign (false) cases of OSF. Mathematically, it is defined as

$$\text{Specificity} = \frac{TN}{FP + TN}\%. \tag{66}$$

Table 1
A 2 × 2 confusion matrix for performance evaluation.

Classifier output	Patients with OSF (as confirmed on biopsy)	
	Negative (absent)	Positive(present)
Negative	TN	FN
Positive	FP	TP

Overall accuracy: The overall accuracy of a test is the measure of "true" findings (true-positive + true-negative results) divided by all test results. This is also termed "the efficiency" of the test.

$$\text{Overall accuracy} = \frac{TP + TN}{TP + FP + FN + TN}\%. \tag{67}$$

7. Results and discussion

The basal cell nuclei boundaries are overlaid on extracted basal layer (shown in Fig. 9) of the H&E image shown in Fig. 5(a). Fig. 7(a–c) shows some of the extracted cells after performing fuzzy classification for identifying cells of NOM and OSF respectively. The segmented nuclei of the cells are shown in Fig. 7(d–f) using GVF. Fig. 10(a) and (c) shows the segmented normal and dysplastic basal cells. Fig. 10(b) and (d) shows the segmented nucleus respectively, which shows the normal nucleus taken very less stain compare to the dysplastic nucleus. This is due to hyperchromatism.

The features are extracted from the segmented basal cell nuclei Fig. 7(d–f). Here we have 771 nuclei for normal and 423 nuclei for OSF with dysplasia.

The features of normal and OSF are summarized into mean, standard deviation (Table 2). The results suggest that 18 features are significant except eccentricity, solidity, rectangularity, orientation and contour irregularity in discriminating normal and OSF group using unsupervised feature selection. An advantage of using the unsupervised feature selection for inspecting feature separability is that the algorithm is generic in nature and has the capability of multiscale representation of the data sets. Fig. 11 shows plot between feature index and feature weights of the unsupervised feature selection between normal, OSF without dysplasia group. Feature weights are basically the distance of k-NN for each feature. Moreover, the plot indicates significance of the feature to discriminate the two groups.

Further, numeric values of most of the feature are increasing steadily from the normal to OSF with dysplasia. The nucleus area of the dysplastic cells is twice as large as that of the normal cells. The increase in nucleus area in this study may be a reflection of the increase in DNA synthesis. The changes occurring in the basal cell nuclei might indicate an increased metabolic activity prior to the invasion of the underlying connective tissues. Thus, the mean intensity of nucleus in sever dysplasia usually appears darker than that of normal nucleus (Fig. 10(b) and (d)), which can be inferred from the results. In normal case the intensity value is 24.98 it indicates stain taken by the nucleus is less but in OSF with dysplasia the intensity value is 18.69, it indicates stain taken by the nucleus is high. This is due to hyperchromatism, i.e., excessive pigmentation in hemoglobin content of basal cell nuclei. It is an important characteristic appearing in a malignant tumor. For the case of sever dysplasia, chromatin abnormality results in increasing staining capacity of nuclei.

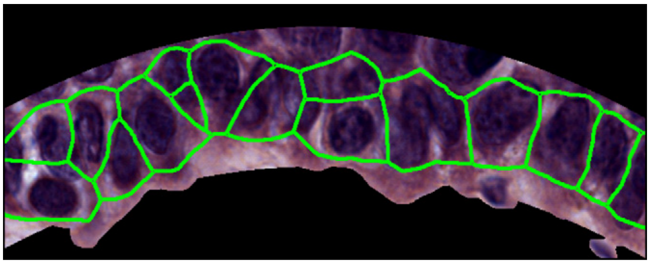

Fig. 9. Segmented boundaries of basal cells are superimposed on the extracted basal layer.

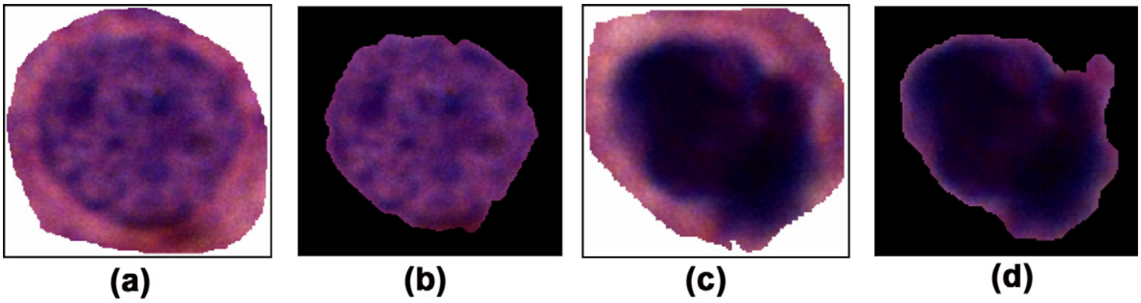

Fig. 10. (a) Segmented normal basal cell; (b) less intense normal basal cell nucleus; (c) segmented dysplastic basal cell; (d) high intense dysplastic basal cell nucleus.

Table 2
Features extracted from nucleus of normal and OSF basal cells.

Sl. no	Nucleus features	Normal μ ± σ	OSF with dysplasia μ ± σ
1	Area	7.93 ± 1.75	13.79 ± 2.07*
2	Perimeter	9.31 ± 1.15	12.50 ± 1.13*
3	Eccentricity	0.89 ± 0.12	0.88 ± 0.14
4	Fourier descriptors	1.01e+013 ± 7.38e+012	8.21e+013 ± 5.17e+013*
5	Zernike moments** (**(*m* = *1*; *n* = *3*))	2.38 ± 0.32	2.44 ± 0.42*
6	Area equivalent diameter	3.16 ± 0.36	4.18 ± 0.31*
7	Perimeter equivalent diameter	2.52 ± 0.56	4.39 ± 0.66*
8	Form factor	1.14 ± 0.08	1.11 ± 0.09*
9	Convex area	8.22 ± 1.82	14.34 ± 2.19*
10	Solidity	0.96 ± 0.01	0.96 ± 0.02
11	Roundness	0.68 ± 0.12	0.70 ± 0.12*
12	Concavity	0.29 ± 0.13	0.55 ± 0.26*
13	Orientation	−1.73 ± 60.21	−1.17 ± 60.09
14	Aspect ratio	10.25 ± 2.26	17.82 ± 2.69*
15	Rectangularity	0.77 ± 0.01	0.77 ± 0.01
16	Area irregularity	1.45 ± 0.69	2.42 ± 1.23*
17	Contour irregularity	0.25 ± 0.07	0.25 ± 0.07
18	Spot areas ratio	1.59 ± 0.07	1.56 ± 0.07*
19	Contrast	0.09 ± 0.03	0.10 ± 0.03*
20	Correlation	0.98 ± 0.01	0.97 ± 0.01*
21	Homogeneity	0.64 ± 0.07	0.59 ± 0.06*
22	Energy	0.91±0.03	0.94 ± 0.02*
23	Mean nuclei intensity	18.69 ± 5.71	24.98 ± 7.03*

[*] Significant based on feature weights.

Fig. 12(a) shows the box plot for one of the feature, area of nucleus, which suggests that median of the feature is almost same as mean so neglecting the chance of outliers for contributing the higher difference between two classes. Fig. 12(b) shows the density plot of perimeter for normal and OSF with dysplasia cases which shows the distinct discrimination between the two groups and 3D scatter plot as shown in Fig. 12(c) shows that the features, i.e., Zernike moments, Fourier descriptors and area equivalent diameter are quiet separable from discrimination point of view with this we can infer a simple linear classifier can achieve higher accuracy.

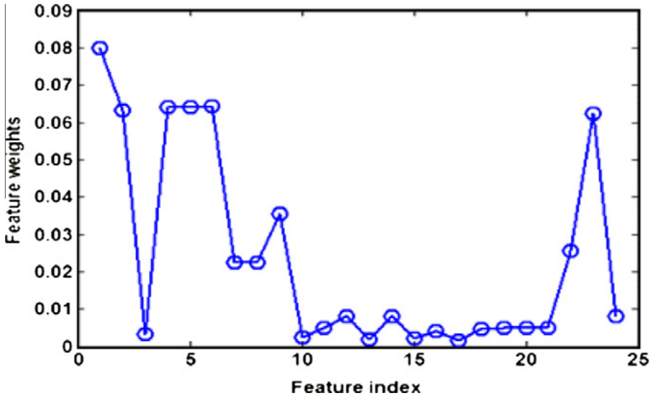

Fig. 11. Plot between feature indexes vs. feature weights for showing significance of features.

We have evaluated the performance of OSF screening system using 341 normal and 429 OSF with dysplasia biopsy images of size 1388 × 1040 pixels obtained from more than 20 patients. To establish the ground truth, biopsy images are commonly graded by a group of experienced pathologists. Before features extraction, nuclei segmentation must be performed. Fig. 6 shows examples of successful nuclei segmentation.

To evaluate the performance of our screening system, we used 1194 nuclei images in this 771 normal nuclei and 423 OSF with dysplasia nuclei images. In our study we have used *k*-fold cross validation for training/testing data partitioning. The advantage of doing this is that we can independently choose how large each test set is and how many trials we average over (Schneider 1997). In our study the number of cases (normal: 771, OSF with dysplasia: 423) is divided by 10 fold; the size of each fold is not the same as shown in Table 3.

Here we have employed two supervised classifiers viz., Bayesian and SVM, three unsupervised classifiers viz., *k*-means, FCM, GMM

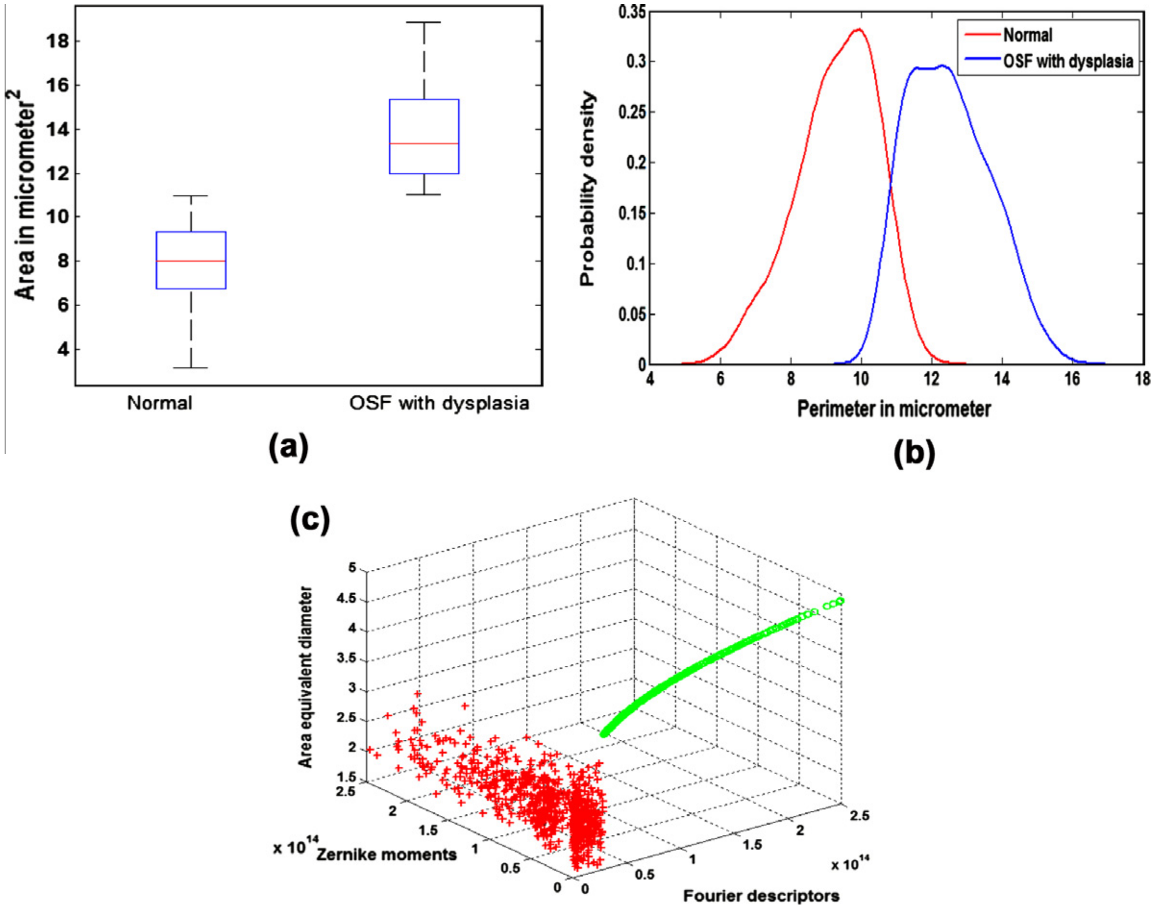

Fig. 12. (a) Box plot for area; (b) Density plot of perimeter for Normal and OSF with dysplasia; (c) 3D plot of features for normal and OSF with dysplasia cases.

Table 3
Stratified 10-fold cross validation of the given data set.

Fold	Size of training set	Size of testing set
Fold#1	1075	119
Fold#2	1074	120
Fold#3	1074	120
Fold#4	1074	120
Fold#5	1074	120
Fold#6	1075	119
Fold#7	1075	119
Fold#8	1075	119
Fold#9	1075	119
Fold#10	1075	119

to evaluate the screening system using 18 features. The best overall performance (99.66%) is obtained with 10-fold cross validation using SVM classifier. The corresponding sensitivity is 99.74% and specificity is 99.53% are also sufficiently high. The supervised classifiers results are listed in Table 4. In case of Bayesian we have obtained 96.56% overall performance. The corresponding sensitivity is 96.43% and specificity is 96.62%. Fig. 13(a–c) shows the sensitivity, specificity and accuracy plot over 10-fold. In SVM we have observed both sensitivity and specificity are more than 99% in all 10-folds consistently, but in Bayesian classifier 7th fold there is a drastic reduction in sensitivity, specificity and accuracy except that all other folds are more than 90%.

Table 4
Performance measure for supervised classifiers.

Classifier	Average sensitivity (%)	Average specificity (%)	Average accuracy (%)
Bayesian	96.43	96.62	96.56
SVM	99.74	99.53	99.66

The classification accuracy is listed in Table 5 for all the three unsupervised classifiers; i.e., *k* means, FCM and GMM classifiers, among them GMM performs well. The best overall performance (90.37%) is obtained using GMM classifier. The corresponding sensitivity is 89.62% and specificity is 91.73% are also sufficiently high. The GMM is trained to classify the data and the log likelihood will converge during estimating model parameters. The log likelihood plot is given in Fig. 14. It converges in seven iterations and becomes stable.

From the above results (Tables 4 and 5), we conclude that the SVM obtains very promising results in classifying the possible OSF patients. We believe that the proposed system can be very helpful to the onco-pathologist for their final decision on their patients. By using such an efficient tool, they can make very accurate decisions.

8. Conclusion

Accurate screening for OSF biopsy images is important to prognosis and treatment planning. Visual grading by human is time-consuming, subjective, and inconsistent while computerized

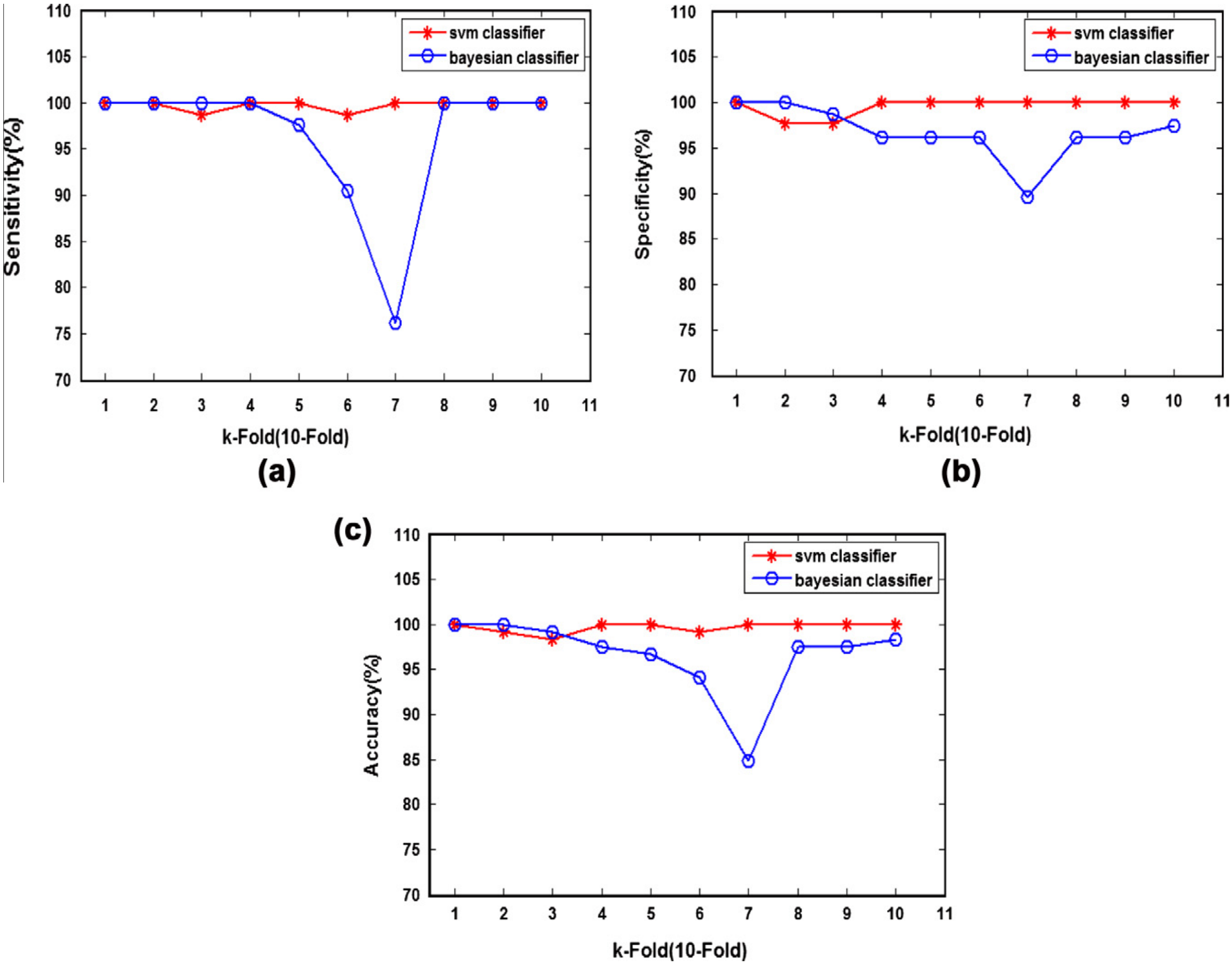

Fig. 13. (a) Sensitivity plot for SVM and Bayesian classifiers over 10-fold; (b) specificity plot for SVM and Bayesian classifiers over 10-fold; (c) accuracy plot for SVM and Bayesian classifiers over 10-fold.

Table 5
Performance measure for unsupervised classifiers.

Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)
k-Means	84.44	83.22	84.00
FCM	90.14	88.18	89.45
GMM	89.62	91.73	90.37

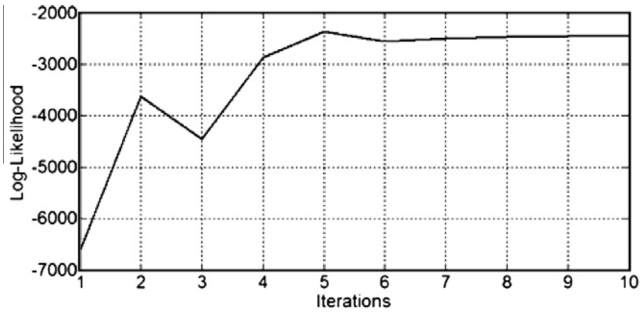

Fig. 14. Log-likelihood values of GMM classifier during training over iterations.

analysis for OSF biopsy images is a very complex task requiring a lot of appropriate image processing steps and experts' domain knowledge for correct screening.

In this paper, we propose an automated system for screening OSF biopsy images. In image preprocessing, a median filtering method is proposed to remove noise. Initially basal layer extracted from histopathological images using various steps viz., fuzzy divergence based thresholding subsequently morphological operations to find the lower boundary of the basal layer and parabola fitting. Further, nuclei are extracted from these cells using color deconvolution, marker-controlled watershed transform and GVF active contour method, such a hybrid approach is robust in terms of removing noise and preserving shapes of nuclei in OSF biopsy images. In feature extraction, 23 features are extracted from segmented biopsy images according to five types of OSF characteristics including nuclear changes (variation in size and shape, polymorphism (nuclei of the basal layer are elongated and perpendicular to basement membrane), nuclear irregularity, hyperchromasia (excessive pigmentation in hemoglobin content of basal cell nuclei) and nuclear texture. These features comprise both local and global characteristics so that normal and OSF with dysplasia can be distinguished effectively. In classification, unsupervised feature selection method is used to select an optimal feature subset (18 features) from the 23 features for the supervised and unsupervised classifiers.

The major contribution of this study is to develop an efficient and effective automated screening system for OSF biopsy images using several methods for image preprocessing, segmentation, feature extraction and image classification. The system is effective because experimental results show that 99.66% of accuracy can be achieved on an average by exercising a set of 341 normal and 429 OSF with dysplasia images obtained from more than 20 patients. A compact set of 18 features and their quantitative measurements are particularly useful for screening is defined in this paper. The best accuracy can be achieved 99.66% using SVM classifier and 90.37% accuracy achieved using GMM classifier because feature subset is carefully selected. We believe that the proposed system can be very helpful to the onco-pathologist for their final decision on to their patients.

Acknowledgement

The authors would like to thank Dr. M. Pal, GNDSIR, Kolkata, India, and Dr. J. Chatterjee, SMST, IIT Kharagpur, India for their clinical support and valuable advices. The authors are very grateful to Mr. Pratik Shah, SMST, IIT Kharagpur, India for assistance during the implementation of the parabola fitting and colour deconvolution algorithms.

References

Banoczy, J. (1982). *Oral leucoplakia* (p. 231). Akademiai Kiado: Budapest.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.

Bilmes, J. A. (1998). *A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models*. Technical Report, UC Berkeley.

Burkhardt, A. (1985). Advanced methods in the evaluation of premalignant lesions and carcinoma of the oral mucosa. *Journal of Oral Pathology, 14*, 751–758.

Chaira, T., & Ray, A. K. (2003). Segmentation using fuzzy divergence. *Pattern Recognition Letters, 24*(12), 1837–1844.

Chaira, T., & Ray, A. K. (2009). *Fuzzy image processing and applications with MATLAB*. New York: CRC Press, pp. 80–81.

Chaudari, D., & Samal, A. (2007). A simple method for fitting of bounding rectangle to closed regions. *Pattern Recognition, 40*, 1981–1989.

Daftary, D. K., Murti, P. R., Bhonsale, R. B., Gupta, P. C., Mehta, F. S., & Pindborg, J. J.(1993). Oral precancerous lesions and conditions of tropical interest. In: Prabhu, S. R., Wilson, D. F., Daftary, D. K., Johnson, N. W., (Eds.), *Oral diseases in the tropics* (pp. 402–424). Oxford: Oxford University Press.

Duda, R., Hart, P., & Stork, D. (2007). *Pattern classification* (2nd ed.). India: Wiley.

Duncan, J. S., & Ayache, N. (2000). Medical image analysis: Progress over two decades and the challenges ahead. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 85–106.

El-Naqa, I., Yang, Y., Wernick, M. N., Galatsanos, N. P., & Nishikawa, M. R. (2002). A support vector machine approach for detection of microcalcifications. *IEEE Transactions on medical imaging, 21*, 1552–1563.

Fan, J., & Xie, W. (1999). Distance measure and induced fuzzy entropy. *Fuzzy Sets Systems, 104*, 305–314.

Farjam, R., Soltanian-Zadeh, H., Zoroofi, R. A., & Jafari-Khouzani, K. (2005). Tree-structured grading of pathological images of prostate. *Proceedings of SPIE: Medical Imaging, 5747,* 840–851.

Fuhrman, S. A., Lasky, L. C., & Limas, C. (1982). Prognostic significance of morphologic parameters in renal cell carcinoma. *American Journal of Surgical Pathology, 6*, 655–663.

Gilles, F. H., Tavare, C. J., Becker, L. E., Burger, P. C., Yates, A. J., Pollack, I. F., et al. (2008). Pathologist interobserver variability of histologic features in childhood brain tumors: Results from the CCG-945 study. *Pediatric and Developmental Pathology, 11*, 08–117.

Glotsos, D. (2003). A hierarchical decision tree classification scheme for brain tumour astrocytoma grading using support vector machines. In *Proceedings of third international symposium on image and signal processing analysis* (Vol. 2, pp. 1034–1038).

Gonzalez, R. C., & Woods, R. E. (2002). *Digital image processing* (2nd ed.). New York: Prentice Hall, pp. 655–659.

Grieg, G., Kubler, O., Kikinis, R., & Jolesz, F. A. (1992). Nonlinear anisotropic filtering of MRI data. *IEEE Transactions on Medical Imaging, 11*(2), 221–232.

Grootscholten, C., Bajema, I. M., Florquin, S., Steenbergen, E. J., Peutz-Kootstra, C. J., Goldschmeding, R., et al. (2008). Interobserver agreement of scoring of histopathological characteristics and classification of lupus nephritis. *Nephrology Dialysis Transplantation, 23*, 223–230.

Hand, J. R., & Broders, A. (1931). Carcinoma of the kidney: The degree of malignancy in relation to factors bearing on prognosis. *Journal of Urology, 28*, 199–216.

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

http://www.cs.cmu.edu/~schneide/tut5/node42.html last accessed March 2010.

http://www.dentistry.bham.ac.uk/landinig/software/software.html last accessed March 2010.

Huang, P. W., & Lai, Y. H. (2010). Effective segmentation and classification for HCC biopsy images. *Pattern Recognition, 43*(4), 1550–1563.

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

Khotanzad, A., & Hong, Y. H. (1990). Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(5), 489–497.

Kim, T. Y., Choi, H. J., Cha, S. J., & Choi, H. K. (2005). Study on texture analysis of renal cell carcinoma nuclei based on the Fuhrman grading system. In *Proceedings of seventh international workshop on enterprise networking and computing in healthcare industry* (pp. 384–387).

Lohse, C. M., Blute, M. L., Zincke, H., Weaver, A. L., & Chenille, J. C. (2002). Comparison of standardized and non-standardized nuclear grade of renal cell carcinoma to predict outcome among 2042 patients. *American Journal of Surgical Pathology, 118*, 877–886.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). Berkeley: University of California Press.

McKeown, M. J., & Ramsey, D. A. (1996). Classification of astrocytomas and malignant astrocytomas by principal component analysis and a neural net. *Journal of Neuropathology and Experimental Neurology, 55*, 1238–1245.

Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(4), 301–312.

Muthu Rama Krishnan, M., Pal, M., Bomminayuni, S. K., Chakraborty, C., Paul, R. R., Chatterjee, J., et al. (2009). Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis-an SVM based approach. *Computers in Biology and Medicine, 39*(12), 1096–1104.

Muthu Rama Krishnan, M., Shah, P., Pal, M., Chakraborty, C., Paul, R. R., Chatterjee, J., et al. (2010). Structural markers for normal oral mucosa and oral sub-mucous fibrosis. *Micron, 41*(4), 312–320.

Muthu Rama Krishnan, M., Pal, M., Paul, R. R., Chakraborty, C., Chatterjee, J., & Ray, A. K. (2010). Computer vision approach to morphometric feature analysis of basal cell nuclei for evaluating malignant potentiality of oral submucous fibrosis. *Journal of Medical Systems*. doi:10.1007/s10916-010-9634-5 [Epub ahead of print].

Novara, G., Martignoni, G., Artibani, W., & Ficarra, V. (2007). Grading systems in renal cell carcinoma. *Journal of Urology, 177*, 430–436.

Paul, R. R., Mukherjee, A., Dutta, P. K., Banerjee, S., Pal, M., Chatterjee, J., et al. (2005). A novel wavelet neural network based pathological stage detection technique for an oral precancerous condition. *Journal of Clinical Pathology, 58*, 932–938.

Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(7), 629–639.

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

Rust, B. W. (2001). Fitting nature's basic functions part I: Polynomials and linear least squares. *Computing in Science and Engineering*, 84–89.

Satheesh, M., Paul, M., & Hammond, S. P. (2007). Modeling epithelial cell behavior and organization. *IEEE Transactions on NanoBioscience, 6*(1), 77–85.

Scarpelli, M., Bartels, P. H., Montironi, R., Galluzzi, C. M., & Thompson, D. (1994). Morphometrically assisted grading of astrocytomas. *Analytical Quantitative Cytology and Histology, 16*, 351–356.

Schad, L. R., Schmitt, H. P., Oberwittler, C., & Lorenz, W. J. (1987). Numerical grading of astrocytomas. *Medical Informatics, 12*, 11–22.

Shabana, A. H., Gel-Labban, N., & Lee, K. W. (1987). Morphometric analysis of basal cell layer in oral premalignant white lesions and squamous cell carcinoma. *Journal of Clinical Pathology, 40*(4), 454–458.

Shuttleworth, J., Todman, A., Norrish. M., & Bennett, M. (2005). Learning histopathological microscopy. *Pattern Recognition and Image Analysis, Pt 2, Proceedings. 3687*, 764–772.

Smith, Y., Zajicek, G., Werman, M., Pizov, G., & Sherman, Y. (1999). Similarity measurement method for the classification of architecturally differentiated images. *Computers and Biomedical Research, 32*, 1–12.

Tabesh, A., Teverovskiy, A. M., Pang, H. Y., Kumar, V. P., Verbel, D., Kotsianti, A., et al. (2007). Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Transaction on Medical Imaging, 26*, 1366–1378.

Vapnik, V. (1998). *Statistical learning theory* (2nd ed.). New York: Wiley.

Xu, C., & Prince, J. L. (1997). Gradient vector flow: A new external force for snakes. In *Proceeding of IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 66–71). Los Alamitos: Comp. Soc. Press.