CHAPTER 2: THEORETICAL BACKGROUND

Presently cancer detection is done by pathologists through evaluation of microscopic biopsy of cancerous tissues by examining the tissue structures, distributions of cells in tissue, regularities of cell shape to determine the level of abnormalities that may be present in the sample under investigation. The outcomes of these examinations may be normal, benign and malignant tissues. The manual evaluation of microscopic biopsy for cancer detection leads to subjective, time consuming and varies with perceptions and level of expertise of pathologists. To overcome these challenges automated cancer diagnosis is needed for objective, fast, accurate and quantitative results. In this chapter, a systematic survey on computational steps for detection of cancer from biopsy images using image processing and pattern recognition tools is presented. These steps involve image pre-processing, enhancement and restoration, segmentation, feature extraction to quantify properties of local area, and classification of sample image into normal and abnormal categories e.g. benign and malignant ones.

2.1 Introduction

Presently cancer is leading cause of the death. About 30% of Indian population gets cancer at some point during their life time. The chances of getting affected by this disease are accelerated due to change in habits in the people such as increase in use of tobacco, deterioration of dietary habits, lack of activities, and many more. The possibility of cure from cancer is increased due to recent combined advancement in medicine and engineering. The chances of curing cancer primarily rely in its early detection and diagnosis. The selection of the treatment

totally depends on its level of malignancy. Medical Professionals use several techniques for detection of cancer. These techniques may include X-Ray, Computer tomography (CT)-Scan, Positron emission tomography (PET), Ultrasound, Magnetic Resonance Imaging) MRI, Urine test, blood test etc. In some cases pathologists uses histopathology biopsy images that is examination of microscopic tissue structure of the patient. Thus biopsy image analysis is a vital technique for cancer detection (Tabesh A. *et al.,* 2007). Histopathology is the study of signs of disease using the microscopic examination of a biopsy or surgical specimen that is processed and fixed onto glass slides. To visualize different components of the tissue under a microscope, the sections are dyed with one or more stains. The aim of staining is to reveal cellular components; counter-stains are used to provide contrast. Hematoxylin-Eosin (H&E) staining has been used by pathologists for over a hundred years. Hematoxylin stains cell nuclei blue, while Eosin stains cytoplasm and connective tissue pink. Due to the long history of H&E, well-established methods, and a tremendous amount of data and publications, there is a strong belief among many pathologists that H&E will continue to be the common practice over the next 50 years (Yang L. *et al.,* 2008). Cytology, on the other hand, is related to the study of cells in terms of structure, function and chemistry. Resulting from the least invasive microscopic biopsies, cytology imagery is the most commonly encountered for both disease screening and biopsy purposes. Additionally, the characteristics of cytology imagery, namely the presence of isolated cells and cell clusters in the images, and the absence of more complicated structures such as glands make it easier to analyze these specimens compared to histopathology.

In histopathology, the cancer detection process normally consists of categorizing the image biopsy in to cancerous one or non-cancerous one. In microscopic biopsy image analysis cytotechnician looks for many of the abnormalities and categorizes the sample based on various characteristics of the cell nuclei such as shape, color, size, proportion to cytoplasm etc.. High resolution Microscopic biopsy provides reliable information for differentiating abnormal and normal tissues.

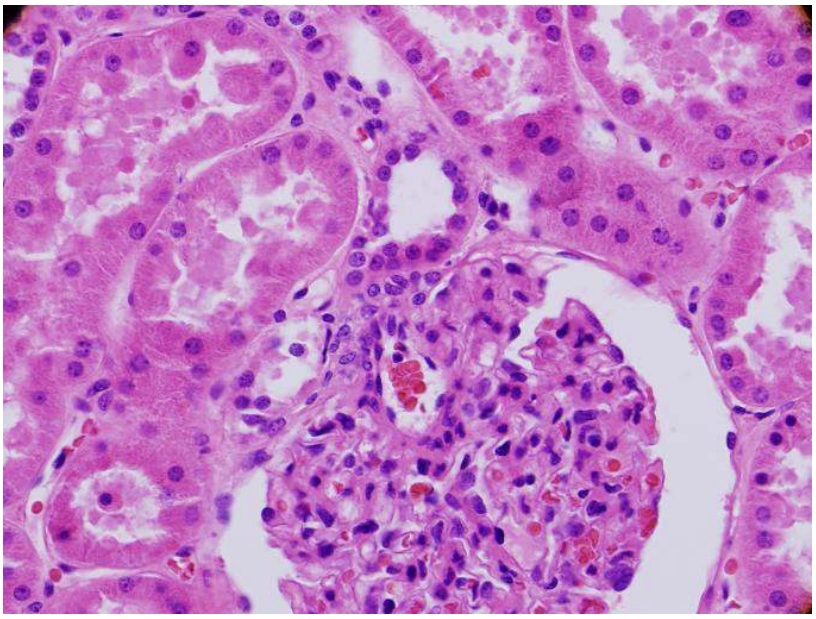

Figure 2.1: Microscopic biopsy image stained with Hematoxylin-Eosin (H&E) (http://www.bioimage.ucsb.edu/images/stories/BioImage/research/Benchmark/BREAST_CANCER/BreastCancerCell_dataset.tar.gz)

The manual examination of microscopic biopsy images may be erroneous and results of the examination may variate for large number of samples due to variations in expertise of the pathologists. Moreover, manual examination of the microscopic biopsy images for cancer detection is subjective and time consuming. Hence, to deal with these limitations the need for the design and development of an automated software tool for the analysis of microscopic biopsy images is

required. Automatic cancer recognition from microscopic biopsy images thus has become an increasingly important task in the medical imaging field (Madabhushi, A. *et al.,* 2009). Some clinical tasks for histopathology image analysis include: Identification of the presence of cancer (image classification); segmenting biopsy images into cancer and non-cancer region (biopsy image segmentation); clustering the tissue region into various classes such as normal, benign, and malignant. Figure 2.1 provides the visual appearance of a microscopic biopsy image.

The automatic analysis system of cell images has to perform a process of segmentation, feature extraction, classification, validation, and error management. Many literature can be found on feature extraction and classification of cell images (Marinakis, Y. *et al.,* 2008), but without the ability of segmenting the nuclei in a robust and accurate way, automation of the whole process is not possible . The biopsy image slides typically contain thousands of cells. Normally, they are scanned with a maximum magnification level of 40×, which results in large images with a size around 80,000 pixels in each dimension. Different magnification levels are used for different tasks, such as background and overall slide quality identification, analysis of cell groupings and clusters, or determination of single cell characteristics.

Metin N. *et al.,* (2009) presented a review on histopathological image analysis for the design of automated cancer detection through a computer aided diagnostic (CAD) system up to year 2009.

In this chapter, a survey for detection of cancer from microscopic biopsy images using image processing and pattern recognition tools are presented up to the year 2014. An integrated framework of CAD tool is examined for microscopic biopsy

images to identify the cancerous tissues. Besides this, step-by-step survey of each procedure used in detection and identification of cancerous and non-cancerous cells from microscopic biopsy images is presented. These analysis procedures are applicable to all imaging modalities for cancer detection from microscopic biopsy images.

2.2 Survey of the Steps involved in the design and development of CAD tool for cancer detection from microscopic biopsy images

The Figure 1.1 shows a framework of computer aided diagnostic (CAD) tool for cancer detection from microscopic biopsy images. There are four fundamental steps involved in design and development of microscopic biopsy images namely Pre-processing, segmentation, feature extraction and finally classification.

The various steps involved in the design and development of the said tool include pre-processing tasks such as restoration and enhancement of microscopic biopsy images for noise reduction and contrast enhancement, segmentation of abnormalities present in the image, features extraction, and classification (Kowala M. *et. al.*, 2013). Surveys of the various design steps for automated CAD system for cancer detection from microscopic images are presented in this work.

2.2.1 Preprocessing

The purpose of the pre-processing is to remove a specific degradation such as noise reduction, contrast enhancement of region of interests etc. The biopsy images acquired from microscope may be defective and deficient in some respect such as poor contrast and uneven staining etc. and they need to be improved through process of image enhancement which increases the contrast between the foreground (objects of interest) and background (Fox .H. *et al.,* 2000). In general, current automatic image- improvement techniques (Persona, P. & Malik, J. 1990)

cannot yet equal what a human operator can achieve interactively. The presence of noise and staining variations within the epithelial layer in sample image necessitates pre-processing.

Major categories of image enhancement and restoration methods include statistical approaches (Black, M. *et al.,* 1998) partial differential equation (PDE)-based approaches (Srivastava, R. *et al.* 2011) wavelet based approaches

(Donoho, D. *et al*., 1995) and many more (Rudin, L. *et al.*, 1992) Traditional enhancement techniques include adaptive filters (Gonzalez, R. C. & Woods, R. E. , 2008).

In the past decades, partial differential equations (PDEs) become an important area of research in image processing. Comparing with the traditional methods of image de-noising PDEs have many advantages, including easy description of local features of an image, employ existing mathematical theory, possible use of many existing numerical algorithms, separate analysis and implementation, preserving most structures and information of an image, deal with the geometric features directly, simulate the dynamic process of image restoration, etc. In the literature, the history of PDE methods dates back to the filter method given by (Lee *et al*., 1980) Based on this research, scale space was introduced by Witkin (Gurcan M. *et al.*, 2009), whereas (Koenderink *et al.,* 2008) made a convolution between image and Gaussian function to implement low-pass filter, which lay a good theoretical foundation of this method. Total variation (TV) (Cataldo, S. Di *et al.*, 2012) methods are very effective for recovering "blocky", possibly discontinuous, images from noisy data. A fixed point algorithm for minimizing a TV penalized least squares functional is presented and compared with existing minimization schemes.

The adaptive filter is more selective than a comparable linear filter, preserving edges and other high-frequency parts of an image. In addition, there are no design tasks; the adaptive filter handles all preliminary computations and implements the filter for an input image. Adaptive filters, however, does require more computation time than linear filtering.

In wavelet based image de-noising and enhancement, at first wavelet transform is applied to the noisy image to produce the noisy wavelet coefficients to the level which we can properly distinguish the occurrence. In second step, appropriate threshold limit is chosen at each level and hard or soft thresholding method is applied to remove the noises. In third step, the inverse wavelet transforms of the thresholded wavelet coefficients is applied to obtain a de-noised image. The above steps can be repeated to number of wavelet transform scales, each representing different degree of wavelet decomposition.

Related to the field of microscopic image processing, (Ruifrok, A. *et al*., 2004) proposed a method for contrast enhancement. In this method the RGB epithelial image is converted to Lab color space and the luminance channel, L is subjected to (a) wiener filtering using a 5×5 filter, (b) gray level shading correction using low pass filtering (Krishnan *et al.*, 2011b) contrast enhancement. The processed L channel is then combined with the two chrominance and converted back to RGB color space. Epithelial cell nucleus absorbs the haemotoxylin while eosin is absorbed by the cytoplasm. The eosin plane is extracted using color de-convolution, thus converting the color image to gray scale image. 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 optical density 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. The length of the vector will be proportional to the amount of stain, while the relative values of the vector describe the actual optical density for the detection channels (M. Gurcan, *et al.* 2012).

The various other methods, such as contrast limited histogram equalization (Wongsritong, K *et al.,* 1998) and many more (Sundareshan, M.K. *et al.*, 1994) also exist in literature which can be utilized for contrast enhancement of microscopic images.

2.2.2 Segmentation

Image segmentation extracts objects/regions of interest from the background; these objects and regions are the focus for further disease recognition and classification. Several promising segmentation techniques are proposed in literature which can reliably overcome histological noise and segment cancer cell nuclei from microscopy biopsy images. Pixel based approaches are the simplest used ones for nuclear segmentation (He L. *et al.,* 2012) They are based on the the information of the pixel value like gray level, color, and texture etc. Rather than pixel based methods the thresholding techniques uses, one or more thresholds that must be determined to satisfy some criteria or to optimize certain objective functions to extract significant objects in biopsy images. Such techniques tend to work only on high-contrast images and do not produce stable results if there is large variability within image sets (Cataldo, *et al.,* 2012) The separation between the different objects (e.g. nuclei, cytoplasm, stroma, and background) is done either by automatic multi-thresholding using image histogram or by pixel classification into groups of

object having similar features (clustering based methods). Other promising techniques are presented recently in the literature providing higher accuracy of separation results than classical methods for various microscopic images.

A combined method based on geodesic active contour and a marker watershed transform was proposed by (Cheng and Sezgin, 1988) First, initial segmentation is obtained using the (Cheng and Sezgin , 1988) model. Then, a marker-controlled watershed transform is applied on the segmented image using a new marking function based on shape markers. A different segmentation approaches are proposed in (Loukas, A. *et al.,* 2004) to automatically separate touching nuclei.

In general, the segmentation approaches can be categorized as follows:

Region based segmentation Methods:

In these methods, image features over an individual region comply with a set of heuristic rules. Simple and straightforward feature thresholding is widely used, due to its computational simplicity and speed, for fast initial segmentation or at intermediate stages of various segmentation scenarios, but usually it cannot stand alone for the final segmentation.

Thresholding approaches:

Thresholding approaches (Gonzalez R.C. 2008) use a value (threshold) to separate objects from background; this value is typically based on image intensity or its transforms such as Fourier descriptors or wavelets. The threshold is usually recognized to satisfy some constraints or to optimize certain objective functions. For example, the generally used Otsu's method finds the threshold to maximize the between-c lass variance (Gonzalez R.C. *et al.,* 2008) For microscopic biopsy image segmentation, multi-thresholding

approaches (M. Sezgin, B *et al.*, 2003) are mostly used to extract objects of different classes, e.g. Nuclei, cytoplasm, stroma, and background.

The multilevel thresholding is necessary to extract different objects from histology images. For example, in the case of K object classes $(s_1, s_2 \ldots\ldots s_K)$ in a digital image I of size $(X \times Y)$, Otsu's method finds the thresholds that maximizes the between-class variance given in equation (2.1).

$$\sigma_B^2 = \sum_{k=1}^{K} p_k (\mu_k - \mu_G) \tag{2.1}$$

where, $p_k = \Sigma_{1 \in S_k} p_1$ (2.2)

p_1 is the normalized histogram (probability) of intensity l, i.e. $p_1 = n_1 / XY$, and n_1 is the number of pixels with intensity l. μ_k is the current mean of s_k, $\mu_k = \left(\frac{1}{p_k} \right) \sum_{i \in S_k} 1 P_1$ (2.3)

and μ_G is the whole image intensity mean. The K classes are separated by $K - 1$ threshold that maximize σ_B^2

Edge based Approaches:

Edge detection (R.C. Gonzalez, 2008) applies as spatial filters (e.g. Prewitt, Canny and Sobel filters) to analyze neighbouring pixel intensity or gradient differences to determine the border among objects and background. Post processing such as edge linking is needed to remove spurious edges and link broken edge segments to form meaningful boundaries.

Watershed approaches:

Watershed methods are frequently used for clustered and touching nuclei separation which is considered as a critical step for segmentation of biopsy images since it has a great impact on breast cancer nuclei quantification. Indeed, classical watershed algorithm suffers from the major drawback of over-segmentation due to the presence of a multitude of regional minima which is typically a consequence of noise. Several techniques are proposed in the literature to reduce the noise sensitivity of the algorithm, such as marker controlled watershed (G. Lin *et al.,* 2003), region merging-watershed, (G. Hamarneh *et al.*, 2009) and watershed method using prior shape (Sahoo, K. *et al*., 1988). All these solutions can partially solve this issue, but they still present many implementation difficulties and require explicit prior knowledge of the image structure.

Contour and shape based approaches: The contour based models are categorized as :

Snakes: Active contours methods (Snakes), which have become powerful tools used for edge detection, medical image segmentation and object tracking (Cataldo *et al.,* 2012) Snakes can be classified into two categories: parametric snakes and geometric active contours (Roehrig, H., *et al.,* 1994). Parametric snakes are explicitly represented as parameterized curves in Lagrange formulation. One shortcoming they have is sensitivity to initialization and lack of ability to handle changes in the topology of the evolving curve. Geometric active contours were introduced more recently and are based on the theory of curve evolution and geometric flows .

Level set: In Active contour the numerical implementation is based on the level set method proposed by (M. Gurcan *et al.,* 2009) and allows segmenting multiple objects automatically at the same time. The classical level set methods (Roehrig, H.*et al.,*1994) need to compute a stopping function based on the gradient of the image, so these models can detect only external boundaries of objects defined by the gradient image. (Chan and Vese 2001), proposed an active contour model without edge (CV model) which is able to detect interior and exterior borders of objects without using an edge function. Some other geodesic active contour models are proposed in (Phukpattaranont P. *et al,*. 2007) which combine the classical active contour model and the two-region segmentation model in order to improve the segmentation accuracy of color images even for discrete and fuzzy edges. None of these models can reliably detect desired objects in the image (i.e. inclusion of other irrelevant objects from the background), and they suffer from slow convergence due to their computational complexity. Moreover, the level set function used in the major active contour formulations is restricted to the separation of two regions. Only few works focus on level set based segmentation in the case of more than two *regions* (X. Long*, et al.,* 2005)

Unsupervised clustering based approaches:

unsupervised techniques, such as K-means, fuzzy c-means, ISODATA clustering, Mean shift (D. Comaniciu, *et al.,* 2002) self-organizing map (Haykin *et al.*, 1999), and adaptive resonance theory (G.A. Carpenter *et al.,* 2003) can be applied to group image points to different objects. In certain

applications such as texture segmentation, feature extraction from the whole slide of biopsy images may be applied before segmentation, which can provide more discriminative features for clustering algorithms than regular intensities and colours.The traditional K-means clustering and a recent nonparametric algorithm, mean shift (G. Hamarneh *et al.,* 2009) clustering. K-means clustering groups image points into *k* clusters by minimizing the objective function given in equation (2.4)

$$\sum_{k=1}^{K} \sum_{i \in S_K} (I_i - \mu_K)^2 \qquad (2.4)$$

where *Ii* is the intensity of the image point x_i in the class s_K Unlike *K*-means clustering, the mean shift algorithm does not assume prior knowledge of the number of clusters. For image segmentation, the image points in a *d*-dimensional ($d = 3$ for colour image) feature space can be characterized by certain probability density function where dense regions correspond to the local maxima (modes) of the underlying distribution. Image points associated to the same mode (by a gradient ascent procedure) are grouped into one cluster. Notionally, the kernel density estimator for n points $(X_i, i = 1, 2, 3 \ldots . n)$ is defined as equation (2.5)

$$f(x) = \frac{1}{n h^d} \sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right) \qquad (2.5)$$

where $K(x)$ is the kernel with the bandwidth h. In practice, radially symmetric kernels such as Epane tichnikov and Gaussian kernels are usually used for clustering. The gradient descent procedure is guaranteed to converge to a point where $\nabla f(x) = 0$, i.e. a local maximum (mode) .

Supervised classification based approaches:

The supervised algorithms that can be applied to build a classifier by a pre-labeled set of training instances. Some of the most well-known methods are artificial neural networks (C.M. Li. *et al.,* 2005) support vector machine (SVM) (Wannous, H. *et al,.* 2007), Bayesian learning (Wang W.W. *et al.*, 2011), and decision trees. Markov random field (MRF) (Monaco, J. *et al.,* 2008), hidden Markov model (HMM) (R.O. Duda, 2000) and conditional random field (CRF) (Lafferty J. *el al.,* 2001) are commonly used to label pixels with the constraints of local pixel distribution, such as Gibbs distribution.

When certain a priori knowledge of image intensity distribution is available, the image segmentation problem can be formulated as a maximum a posterior estimation with Bayesian model. In principle, given the observed image I, the objective is to label image pixels to different classes $\{S = (s_1, s_2, \ldots\ldots s_K)\}$, which maximize the posterior probability of the labeling configuration F given the observation I in equation (2.6) as;

$$P(F \setminus I) = \frac{P(I \setminus F)P(F)}{P(I)} \alpha\, P(I \setminus F)P(F) \tag{2.6}$$

Because the density function $P(I)$ is a factor to ensure the total probability is one, it can be omitted so that $P(F \setminus I)$ is proportional to the product of the likelihood $P(I \setminus F)$) and the prior probability $P(F)$. The likelihood function characterizes the intensity distributions in different image regions, e.g. Gaussian distributions calculated by equation (2.7) as:

$$P(I \setminus F = S_k) = N(I \setminus K) \tag{2.7}$$

K is the mean and variance of the class s_K. With MRF models, the prior probability $P(F)$ is constructed on a small neighborhood.

Hybrid segmentation approaches:

The combination of above approaches considered as hybrid approaches. Combining a priori boundary, shape, region and feature information of the cancerous gland improves segmentation accuracy. Such methods are robust to noise and produce superior results in the presence of shape and texture variations of the prostate.

A brief survey of various segmentation approaches for microscopic biopsy images recently reported in literature is presented below.

In work (Devrim O. *et al.,* 2012), presented an automatic system for segmenting HCC biopsy images. In preprocessing, they used a dual morphological gray scale reconstruction method to remove noise and accentuate nuclear shapes. A marker-controlled watershed transform was applied to obtain the initial contours of nuclei and a snake model was used to segment the shapes of nuclei smoothly and precisely. Finally, a SVM-based decision-graph classifier to classify HCC biopsy images was proposed. Experimental results showed that 94.54% of classification accuracy could be achieved by using the SVM-based decision-graph classifier while 90.07% and 92.88% of classification accuracy could be achieved by using k-NN and SVM classifiers, respectively.

Yousef *et al.,* (2010) presented a robust and accurate method for segmenting cell nuclei using combination of ideas. The image foreground was extracted automatically using graph-cut based binarization. Then by combining multicsale Laplacian of Gaussian (LoG) the nuclear seed points were detected. These points

were used to perform initial segmentation and again refined using second graph-cut algorithms. It used 25 biopsy images with 7400 nuclei and four segmentation errors were evaluated. The overall accuracy obtained by this algorithm was about 86%. Authors presented two automated methods for the segmentation of histology tissue images that overcome the limitations of the manual approach as well as of the existing computerized techniques. In this work, the first independent method, based on unsupervised color clustering, automatically recognizes the target cancerous areas in the biopsy specimen and disregards the stroma; the second method, based on colors separation and morphological processing, exploits automated segmentation of the nuclear membranes of the cancerous cells. The experimental results on real tissue images demonstrated the accuracy of the techniques compared to manual segmentations; additional experiments show that our techniques are more effective in histology images than popular approaches based on supervised learning or active contours. The presented procedure can be exploited for any applications that require tissues and cells exploration and to perform reliable and standardized measures of the activity of specific proteins involved in multi-factorial genetic pathologies.

Cataldo S. Di *et al.,* (2012) presented a novel synergistic boundary and region-based active contour model that incorporates shape priors in a level set formulation with automated initialization based on watershed. An application of these synergistic active contour models using multiple level sets to segment nuclear and glandular structures on digitized histopathology images of breast and prostate biopsy specimens are demonstrated in this work. Unlike previous related approaches, this model is able to resolve object overlap and separate occluded boundaries of multiple objects simultaneously. The energy functional of the active

contour is comprised of three terms. The first term is the prior shape term, modeled on the object of interest, thereby constraining the deformation achievable by the active contour. The second term, a boundary-based term detects object boundaries from image gradients. The third term drives the shape prior and the contour towards the object boundary based on region statistics. The results of qualitative and quantitative evaluation on 100 prostate and 14 breast cancer histology images for the task of detecting and segmenting nuclei and lymphocytes reveals that the model easily outperforms two state of the art segmentation schemes (geodesic active contour and Rousson shape-based model) and on average is able to resolve up to 91% of overlapping/occluded structures in the images.

Sahirzeeshan Ali *et al.,* (2012) presented an unsupervised approach for the segmentation and classification of cells. The segmentation process involve automatic thresholding to separate the cell regions from the back ground, a multi scale hierarchical segmentation algorithm to partition these regions based on homogeneity and circularity, and a binary classifier to finalize the separation of nuclei from cytoplasm within the cell regions. Classification is posed as a grouping problem by ranking the cells based on their feature characteristics modeling the degree of abnormalities. The presented procedure constructs a tree using hierarchical clustering, and then arranges the cells in a linear order by using an optimal leaf ordering algorithm that maximizes the similarity of adjacent leaves without any requirement for training examples or parameter adjustment. Performance evaluation using two data sets show the effectiveness of the proposed approach in images having inconsistent staining, poor contrast, and overlapping cell.

Aymen Mouelhia *et al.,* (2013) presented an algorithm for semi-automatic segmentation of the nuclei under adequate control of the expert user. It can work automatically or interactively guided, to allow for segmentation within the whole range of slide and image characteristics. It facilitates data storage and interaction of technical and medical experts, especially with its web-based architecture. The algorithm localizes cell nuclei using a voting scheme and prior knowledge, before it determines the exact shape of the nuclei by means of an elastic segmentation algorithm. After noise removal with a mean-shift and a median filtering takes place, edges are extracted with a canny edge detection algorithm. Motivated by the observation that cell nuclei are surrounded by cytoplasm and their shape is roughly elliptical, edges adjacent to the background are removed. A randomized Hough transform for ellipses finds candidate nuclei, which are then processed by a level set algorithm. The algorithm is tested and compared to other algorithms on a database containing 207 images acquired from two different microscope slides, with promising results

Saima Rathore *et al.,* (2014) presented a new automatic system to perform both segmentation of touching nuclei in order to get the total number of cancer nuclei in each class. In this work, a modified geometric active contour model is used for multiple contour detection of positive and negative nuclear staining in the microscopic image. Then, a touching nuclei method based on watershed algorithm and concave vertex graph was proposed to perform accurate quantification of the different stains. Finally, benign and malignent nuclei were identified by their morphological features and they were removed automatically from the segmented image for positive cancer nuclei assessment. The segmentation algorithms are tested on two datasets of breast cancer cell images containing different level of

malignancy. The experimental results showed the superiority of the proposed methods when compared with other existing segmentation methods. On the complete image database, the segmentation accuracy in term of cancer nuclei number is over than 97%, reaching an improvement of 3–4% over earlier methods.

Metin N. Gurcan *et al.,* (2009) presented a computer-aided design for automatic cell segmentation and nucleus-to-cytoplasm (NC) ratio analysis. The experimental results show that the method presented in this work provides objective segmentation results with high efficiency and consistent accuracy. In addition, the evaluated NC ratio values are very close to the results of manual cell segmentation, indicating that the proposed work in this work consider not only the performance of analysis procedure but also practical criteria as well as clinical requirement which has significant potential for biomedical imaging analysis and medical values in a variety of applications. It can help medical doctors to noninvasively and immediately identify early symptoms of diseases, especially fatal diseases like cancer, that involve abnormal NC ratios. Moreover, the determination of NC ratios of skin cells using the proposed automatic algorithm is more objective and robust than that using manual approach, and hence medical doctors can diagnose potential diseases without the influence of any subjective factor, such as the subjective judgment of analyzer and the fatigue of the medical personnel. It can also quantify several physical factors correlated with NC ratios or cell.

S. Di Cataldo *et al.,* (2009) presented a new learning method, multiple clustered instance learning (MCIL), for histopathology image segmentation. An integrated framework to classify histopathology images as having cancerous regions or not,

segment cancer tissues from a cancer image, and cluster them into different types were developed. The system reported in this work automatically learns the models from weakly supervised histopathology images using multiple clustered instance learning (MCIL), derived from MIL. Many previous MIL-based approaches have achieved encouraging results in the medical domain such as major adverse cardiac event (MACE) prediction (S. Di Cataldo *et al.,* 2009) .The classification (cancer vs. non-cancer image), medical image segmentation (cancer vs. non-cancer tissue), and patch-level clustering (different classes) are used . It embed the clustering concept into the multiple instance learning (MIL) setting and derive a principled solution to performing the above three tasks in an integrated framework. In addition, the work introduced a contextual constraint as a prior for MCIL, which further reduces the ambiguity in MIL. Experimental results showed on histopathology colon cancer images and cytology images demonstrated the great advantage of MCIL over the competing methods and the accuracy was 98.92%.

Madhumala Ghosha *et al.,* (2010) introduced an automated approach to leukocyte recognition using fuzzy divergence and modified thresholding techniques. The recognition is done through the segmentation of nuclei where Gamma, Gaussian and Cauchy type of fuzzy membership functions are studied for the image pixels. It is in fact found that Cauchy leads better segmentation as compared to others. In addition, image thresholding is modified for better recognition.

The Table 2.1 lists a brief introduction about microscopic biopsy image segmentation approaches and data set used for segmentations.

Table 2.1: Previous works reported in the literature for the segmentation of microscopic biopsy images

Authors (Year)	**Methods used for Segmentation**	**Objects for segmentation**	**Data set used**
Po-.Whei Huang *et al*., (2009)	Marker control watershed transform	Nucleus and cytoplasm	1000×1000,4000×3000 and 275×275 HCC biopsy images
Metin N. *et al.,* (2009)	Shaped based Approaches, Active contour , fuzzy c-means, and watershed method	Cell Nuclei. Cell detection rate achieved about 90-92%	20 biopsy of size 512×512×512 of 4 billion pixels
Yousef Al-Kofahi *et al*., (2010)	Graph cut methods	Cell Nuclei. The overall accuracy of segmentation algorithms was 92.6%.	15 biopsy images with 74000 nuclei
He Lei *et al* ., (2012)	Thresholding, active contours, and Markov random fields (MRF)	Nuclei, cytoplasm, stroma	Digitized histology images
Christoph *et al*. , (2012)	Mean shift filtering,	Cell and cytoplasm	Microscopic images

	thresholding	segmentation	
A.C. Cen *et al.,* (2012)	Thresholding, watershed, and active contours	Nuclei, area, radix index with 94% accuracy	Microscopic biopsy images
Devrim *et al.,* (2013)	Texture based segmentation	Cell carcinoma detection with 95.5% accuracy	230 light microscopic images 4080×3720 images 256 bit /channel
Kowal *et al*., (2013)	Adaptive thresholding, k-means, fuzzy c-means (FCM) and k-nearest neighbour (KNN)	Nuclei and cytoplasm with 96% segmentation accuracy	704×578, BMP image files with 8 bit/channel RGB.
Aymen Mouelhi *et al*., (2013)	Geometric active contour, water shed	Nuclei touching and nuclei cytoplasm segmentation with accuracy 96.5%	2048×1360 JPEG files with 24 bit/channel
Xu.Y. *et al.,* (2014)	Conditional random field (CRF) based approach	Nuclei segmentation with 98% accuracy	60 RGB biopsy images
Yan Xu, *et al.,*	Multiple clustered	Segmentation of	Sample biopsy

(2014)	instance learning (MCIL)	cancerous and non-cancerous tissue with 96% accuracy	images
Madhumala Ghosha *et al.,* (2010)	Fuzzy divergence and modified thresholding	Segmentation of cancerous and non-cancerous tissue	Better segmentation

2.2.3 Feature Extraction

After segmentation, feature Extraction is one of the important steps in analysis of biopsy images. Microscopic biopsy images features can be extracted at tissue level or cell level. The shapes of nuclei are no longer kept round in cancerous tissues because of their serious deformity. For better capturing the shape information, we use both region-based and contour-based methods to extract anti-circularity, area irregularity, and contour irregularity of nuclei as the three shape features to reflect the irregularity of nuclei in biopsy images. The cellular-level features focuses on quantifying the properties of individual cells without considering spatial dependency between them. In microscopic biopsy images for a single cell, the morphological, textural, fractal, and/or intensity-based features can be extracted. The tissue level features quantify the distribution of the cells across the tissue; for that, it primarily makes use of either the spatial dependency of the cells or the gray-level dependency of the pixels. For a tissue, the textural, fractal, and/or topological features can be extracted. We are concentrating few features of microscopic biopsy images that are very useful in pathological interpretation.

There are five interesting characteristics of microscopic biopsy images. These characteristics are: nuclear size, nucleocytoplasmic ratio, nuclear irregularity, hyperchromatism, and anisonucleosis. Based on these characteristics, many important features can be extracted from microscopic biopsy images.

In addition to the above mentioned statistical and shape based features, the texture and spectral features of microscopic images can also be extracted for clinical significance. These features may be extracted using various methods such as gray level co-occurrence matrix (GLCM), linear binary patterns (LBP), and Laws texture energy (LTE), and higher order spectra (HOS) etc. Some significant works had been contributed by various researchers for microscopic biopsy image analysis that contains variety of features to derive clinically significant information. Most of the researchers has used different texture properties along with geometric features for the identification of cancer from microscopic biopsy images .

Mohapatra *et al.,* (2011) used two types of features i.e. texture features using gray level co-occurrence matrices (GLCM) (Sonka *et al*., 2007) and shape features to automate the cancer detection in blood smear images. Further, (Sinha and Ramakrishnan *et al.,* (2003)) also used shape, color and texture to differentiate the WBC in blood smear images. Rezatofighi and Soltanian-Zadeh (2011) extracted some morphological features like nucleus area and perimeter, number of the separated parts of nucleus, etc. With color features and textural features. Texture features are calculated using co-occurrence matrix and the local binary pattern. Further, Kuse *et al*., (2010) has used eighteen texture features derived from the gray level co-occurrence matrix for the identification of cancer in the H&E stained histopathological

images. Moreover, a large set of features have been obtained by Osowski *et al.,* (2009) for the recognition of blood cells using genetic algorithm (GA) and a support vector machine (SVM).

Traditional features (Zhang, B., Gao, Y., Zhao, S., Liu, J., 2010.) include morphometric with object size and shape (e.g. compactness and regularities), graph- based features (e.g. Voronoi diagrams, Delaunay triangulation, and minimum spanning trees), intensity and color features (e.g. statistics in different color spaces); and texture features (Haralick R.M., 1973) entropy, Gabor filter, power spectrum, co-occurrence matrices, and wavelets). In addition, besides using the biopsy image in the spatial domain, many features can also be extracted from the other transformed spaces, like Frequency (Fourier and wavelet transforms). With multiple classes of features extracted from large size biopsy images, the resulting vast quantity of data (e.g. a feature vector with thousands of elements for each pixel) can be prohibitive for feasible analysis, even with current high performance computing machines. Therefore, feature dimensionality reduction (DR) are needed for select the most discriminative features. Commonly used DR tools include both linear and nonlinear techniques. Linear techniques, such as principal component analysis (PCA), linear discriminant analysis (LDA) (Petrou, M., Sevilla, P.G., 2006) and multidimensional scaling (MDS) are used in case of linearly separable points in the feature space. Principal component analysis, linear discriminant analysis and multi-dimensional analysis, feature extraction techniques assume Euclidean distance among the feature points.

As an unsupervised data analysis tool, PCA finds orthogonal eigenvectors (i.e., principal components) along which the greatest amount of variability in the

data lies. However, the projection of feature points to the principal component directions may not separate the data well for classification. In contrast, LDA is a supervised learning tool, which incorporates data label information to find the projections that maximize the ratio of between-classes variance and within-classes variance. MDS, on the other hand, reduces data dimensionality by preserving the least squares Euclidean distance in low-dimensional space. For nonlinear DR techniques such as spectral clustering (Yu-Li You, *et al*., 2000), isometric mapping Isomap (Leonid I. Rudin, *et al.,*2009) locally linear embedding (LLE) (B.W. Matthews, *et al.,* 1973) and Laplacian eigenmaps (LEM) (Lee J. S. *et al.,* 2005) Euclidean relationship among the feature points is not assumed. Above mentioned techniques are more suitable for inherently nonlinear biomedical structures. Spectral clustering algorithms (graph embedding) employ graph theory to partition the graph into clusters and separate them accordingly. The ISO map algorithm estimates geodesic distances among points along the manifold (i.e., nonlinear surface embedded in high-dimensional space along which dissimilarities between data points are best represented), and preserves the nonlinear geodesic distances (as opposed to Euclidean distances used in linear methods) while projecting the data onto a low-dimensional space. LLE uses weights to preserve local geometry in order to find the global nonlinear manifold structure of the data. Similar to spectral clustering, Isomap, and LLE, the LEM algorithm makes local connections but utilizes the Laplacian to simplify the determination of the locality preserving weights, which are used to obtain the low-dimensional data embedding. Based on the simplified feature vectors obtained by dimensionality Reduction techniques, classification algorithms can be applied to

identify diseases by comparing the input image features with a set of pre-derived training sample features.

L. Jiang, W. Yang, (2003) presented a feature descriptor which utilizes fractal geometric analysis with four multi-fractal measures to construct an eight dimensional feature space. Bag-of-feature-based classification model proposed to discriminate a set of hepatocellular carcinoma images into five categories according to Edmondson and Steiner's grading system. Three feature selection methods were utilized to obtain the most discriminative features of code word dictionary. Furthermore four other textural feature descriptors: Gabor-filters, LM-filters, local binary patterns, and Haralick, to obtain a benchmark of the accuracy of the classification are incorporated. Experimental results indicated the significance of the multifractal features for describing the histopathological image texture because it outperformed other four feature descriptors.

Christoph Bergmeir *et al.,* (2012) presented model to extract the texture features by using local histograms and co-occurrence matrices. The quasi-supervised learning algorithm operates on two datasets, one containing samples of normal tissues labeled only indirectly, and the other containing an unlabeled collection of samples of both normal and cancer tissues. In this framework several texture feature vector datasets corresponding to different extraction parameters were tested . The Independent Component Analysis dimensionality reduction approach was also identified as the one improving the labeling performance evaluated in this series. In this work 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 labeled ground-truth datasets in a quasi-supervised strategy. The resulting labeling performances were compared to that of a conventional powerful supervised classifier using manually labeled ground-truth data. The supervised classifier results were calculated false positive and true positive rate as 3.5% and 95% for the same case.

(Haralicks R.M. *et al.,* 1973) presented histogram of oriented gradients (HOG), and Color component based statistical moments (CCSM), features selection and extraction approaches to classify the cancerous cells from microscopic biopsy images the feature extracted by authors are Contrast ,correlation, energy, homogeneity, (Haralicks M *et al*. 2012) RGB, Gray Level, HSV. The results were tested on 174 colon biopsy images and improved performance calculated as 98.85%.

M. Muthu Rama Krishnan et *al,.* (2012) introduced methodology for feature extraction and classification of histology images into normal, OSF without Dysplasia (OSFWD) and OSF with Dysplasia (OSFD), which would help the oral onco-pathologists to screen the cancerous portion rapidly. In this work the optical density of the pixels in the light microscopic images is recorded and represented as matrix quantized as integers from 0 to 255, for each fundamental color (Red, Green, Blue), resulting in a M×N×3 matrix of integers. Depending on either normal or OSF condition, the image has various granular structures which are self-similar patterns at different scales termed "texture". They have extracted these textural changes using Higher Order Spectra (HOS), Local Binary Pattern (LBP), and Laws Texture Energy (LTE) from the histopathological images (normal, OSFWD and OSFD). These feature vectors were fed to five different

classifiers: Decision Tree (DT), Sugeno Fuzzy, Gaussian Mixture Model (GMM), *K*-Nearest Neighbor (*K*-NN), Radial Basis Probabilistic Neural Network (RBPNN) to select the best classifier. Our results show that combination of texture and HOS features coupled with fuzzy classifier resulted in 95.7% accuracy, sensitivity and specificity of 94.5% and 98.8% respectively.

Landini G *et al.,* (2009) a method for morphologic characterization of cell neighborhoods in neoplastic and preneoplastic epithelium is presented. The major objective of the work was to explore tissue organization of cell neighborhoods in histologic preparations. The local complexity of solid tissues was measured in images of discrete tissue compartments. The exclusive areas associated with cell nuclei (v-cells) were computed using a watershed transform of the nuclear staining intensity. Mathematical morphology was used to define neighborhood membership, distances and identify complete nested neighborhoods. Neighborhood complexity was estimated as the scaling of the number of neighbors relative to reference v-cells. The methodology applied to hematoxylin-eosin-stained sections from normal, dysplastic and neoplastic oral epithelium revealed that the scaling exponent, over a finite range of neighborhood levels, is nonunique and fractional. While scaling values overlapped across classes, the average was marginally higher in neoplastic than in dysplastic and normal epithelia. The best classificatory power of the exponent was 58% correct classification into 3 diagnostic classes (11 levels) and 83% between dysplastic and neoplastic classes (13 levels).V-cell architecture retains features of the original tissue classes and demonstrates an increase in tissue disorder in neoplasia. This methodology seems suitable for extracting information from tissues where identification of cell boundaries (and therefore segmentation into individual cells) is unfeasible

Rasha Abu *et al.,* (2011) presented a number of innovative techniques to assess a number of morphological features of different grades of oral epithelial dysplasia. It was observed that the epithelial lining of the oral cavity can sometimes experience certain changes that put it at a higher risk of undergoing malignant transformation. Such changes present clinically as 'premalignant' lesions that at the histological level feature pathological alterations known as epithelial dysplasia. However, the degree of alteration of tissues is routinely assessed visually, thus introducing an element of subjectivity to the diagnostic process. The aim of this work was to apply objective and quantitative image analysis techniques to one problematic area in histopathological diagnosis i.e. the grading of the severity of epithelial dysplasia. Histopathological diagnosis, which depends to some degree on individual judgment of histological features by an observer, has been shown to be subject to intra-and inter-observer variations that affect the accuracy and reproducibility of the diagnostic process.

M. Muthu Rama Krishnan *et al.,* (2012) introduced 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. In experimental results, basal cells constitute the proliferative compartment (called basal layer) of the epithelium for the 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 authors proposed to initially extract the basal layer 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 de-convolution, marker-controlled watershed transform and gradient vector flow (GVF) active contour method. Eighteen morphological, 4 gray-level co-occurrence matrixes (GLCM) based texture features and 1 intensity feature was quantized from five types of basal nuclei characteristics.

Table 2.2 illustrates the contributions of feature extraction algorithms, by various researchers for microscopic biopsy images.

Table 2.2: Previous work reported in the literature for the feature extraction from microscopic biopsy images

Authors (Year)	**Methods used for Feature Extraction and Selection.**	**Feature Extracted**	**Data set used and performance measures**
Anant Madhbhushi *et. al*., (2007)	Graph- based features extraction	Shape, size, center of mass, texture feature, and spatial related features	Number of nodes, number of edges , sensitivity and accuracy etc.
Metin N. Gurcan *et al.,* (2009)	Graph- based features extraction	Size, shape, statistical and texture features	Number of cells, number of triangles , and N/C
M.M.R. Krishnan *et al*., (2012)	A hybrid feature extraction (LBP, LTE, and HOS) paradigm is used for feature extraction	Normal, OSFWD and OSFD identified for cancer detection from histopathology images	TP ,TN, FP and FN, Non-cancerous and OSFWD and OSFD were detected.
A.D. Belsare *et al,* . (2012)	Texture, graph, morphological, and Voroni diagram features.	Smoothness, coarseness, regularity, correlation, contrast, number of nuclei, shape size and roundness	Normal, abnormal, and grade of cancer.

L. He. *et al. ,* (2012)	Morphometry,topological intensity,color texture	Area, size, boundary, shape, moments, Harallick's and Gabor texture features, Markovian, run lenth texture, wavelet density features etc.	TP ,TN, FP and FN
Onder *et al.,* *(*2013)	Texture feature,	Variance, kurtosis , mean value of pixels, entropy , energy , contrast, co-relation etc.	Ground truth atlas data, ROC etc.
S. Rathore *et al. ,* (2014)	Harlicks ,histogram of oriented gradients, component based statistical moments	Contrast, correlation, energy, homogeneity, Harallick's texture features, RGB, gray Level, and HSV features.	Tested on 174 colon biopsy images and achieved accuracy of 98.85%.
Landini G *et al. ,* (2010)	Explore tissue organization of cell neighborhoods in histologic preparations	Different grades of oral epithelial dysplasia based features	Achieved correct classification of 58% into 3 diagnostic classes (11 levels) and 83% correct classification between dysplastic tissue
Rasha Abu Eid *et al.,* (2011)	Innovative techniques to assess a number of morphological features	Different grades of oral epithelial dysplasia based features	Morphological features were extracted
M. Muthu Rama Krishnan *et al.* ,	Texture morphological	Eighteen morphological, 4 gray-level co-	OSF is selected from NOM

(2012)	and intensity based features were extracted	occurrence matrix (GLCM) based texture features and one intensity feature	

2.2.4 Classification

After feature extraction, the classification is another challenging task for automatic detection of cancer from microscopic biopsy images. Classification should give the answer whether microscopic biopsy is benign or malignant. For classification purposes, many classifiers have been used. Some commonly used classification methods are: artificial neural networks (ANN), Bayesian classification, K-nearest neighbour classifiers, support vector machine (SVM) and decision trees (DTs).

Here few recently used classifiers for microscopic biopsy image classification methods are described.

Saima Rathore *et al,.* (2014) presented colon biopsy image classification (CBIC). Authors tested 174 colon biopsy to classify with linear, radial basis function (RBF) and sigmoid SVM and achieved the performance of classification about 98.85%.

Metin N, *et al.,* (2009) presented classification of histology images for four carcinoma types: cervix, prostate, breast and lung. In this work, for cervix carcinoma a Bayesian belief network is used to construct decision support system for automatic determination of grade of cervical intraepithelial neoplasia (CIN). SVM and MRF are used to classify and grade the prostate carcinoma from prostate histology image. The breast carcinoma is classified by Bloom-Richardson(BR) grading scheme. Dimensionality reduction is done by the principle component analysis and SVM is used to classify the breast carcinoma.

For lung carcinoma LDA is used to extract features and SVM is used to classify the lung carcinoma in to small cell carcinoma and non-small cell carcinoma.

Manjunath B.S. *et al.,* (1996) presented a novel methodology for automatic clinical prediction of renal tumor cancer from histology images using shape-based features. These shape-based features describe the distribution of shapes extracted from three dominant H&E stain colors in renal tumor histopathological images. Authors evaluated the four-class prediction performance of shape-based classification models using 10 iterations of three-fold nested CV. The overall classification accuracy of 77% (average external CV accuracy) is favourable compared to previous methods that use traditional textural, morphological, and wavelet-based features. Moreover, results indicate that combining shape-based features with traditional histological image features can improve prediction performance. The biological significance of the characteristic shapes identified by this algorithm suggests that this automatic diagnostic system mimics the diagnostic criteria of pathologists.

Juan C. Caicedo *et al,.* (2009) has extracted features of 1502 histology images with 18 different concepts. The classification strategy is based on binary classifiers following the one-against-all rule. For each experiment, the regularization parameter of the SVM is controlled by using 10-fold cross validation in the training dataset, to guarantee good generalization on the test dataset. Reported results are calculated on the test dataset and averaged over all 18 classes.

Krishan M. M. R. *et al.,* (2011) presented a comparative study of automated diagnosis of oral cancer using higher order spectra features and local binary pattern features extracted from the epithelial layer in classifying normal, OSFWD

and OSFD. Authors improved the classification accuracy based on textural features for the development of a computer assisted screening of oral sub-mucous fibrosis (OSF). The approach introduced was used to grade the histopathological tissue sections into normal, OSF without dysplasia (OSFWD) and OSF with dysplasia (OSFD), which would help the oral onco-pathologists to screen the subjects rapidly. For this purpose, they extracted twenty three HOS features and nine LBP features and fed them to a Support Vector Machine (SVM) for automated diagnosis. One hundred and fifty eight images (90 normal, 42 OSFWD and 26 OSFD images) were used for analysis. LBP features provide a good sensitivity of 82.85% and specificity of 87.84%, and the HOS features provide higher values of sensitivity (94.07%) and specificity (93.33%) using SVM classifier.

Muthu Rama Krishnan *et al.,* (2011b) presented a quantitative microscopic approach for discriminating inflammatory and fibroblast cells of oral submucous fibrosis (OSF) from normal oral mucosa (NOM) in respect to shape features of the sub-epithelial connective tissue (SECT) cells. They also used segmentation and classification of sub-epithelial connective tissue (SECT) cells except endothelial cells in oral mucosa of normal and OSF conditions has been reported. Segmentation has been carried out by colour deconvolution and subsequently the cell population has been classified using Support Vector Machine (SVM) based classifier. Moreover, the shape features used in this study were statistically significant using Mann Whitney U test, which enhance the statistical learning potential and classification accuracy of the classifier. Automated classification of SECT cells characterizes this precancerous condition very precisely in a

quantitative manner and unveils the opportunity to understand OSF related changes in cell population having definite geometric properties.

Mookiah MR *et al.,* (2011) presented an automated diagnostic methodology based on textural features of the oral mucosal epithelium to discriminate normal and oral submucous fibrosis (OSF). A total of 83 normal and 29 OSF images from histopathology sections of the oral mucosa were considered. The diagnostic mechanism reported in this work consists of two parts: feature extraction using Brownian motion curve (BMC) and design of a suitable classifier. The discrimination ability of the features has been substantiated by statistical tests. An error back-propagation neural network (BPNN) is used to classify OSF vs. normal. Fisher's linear discriminant analysis yields 100% sensitivity and 85% specificity, whereas BPNN leads to 92.31% sensitivity and 100% specificity, respectively.

M. Muthu Rama Krishnana *et al.,* (2011a) authors presented and discussed the approaches for textural characterization of histopathological images for oral sub-mucous fibrosis detection. The aim of the work presented in this work was to improve the classification accuracy based on textural features for the development of computer assisted screening of oral sub-mucous fibrosis (OSF). In this work authors extracted 71 textural features from epithelial region of the tissue section using various wavelet families, Gabor wavelets, local binary patterns (LBP), fractal dimension and Brownian motion curve. SVM classifier was used for classification purposes and accuracy of 88.38% was achieved.

Krishnan, Muthu Rama *et al.,* (2011) developed a knowledge-based segmentation algorithm using anisotropic diffusion and fuzzy divergence based thresholding followed by colour based region growing. They extracted the mean thickness of the Basement membrane (BM). The significance of the extracted feature (thickness) was evaluated using statistical analysis and it showed that the feature was significant in discriminating the three groups. Further, they also observed that there is an increasing trend of BM thickness for OSFWD and OSFD compared to normal counterpart. The significant features were fed to the support vector machine (SVM) classifier to discriminate (classify) normal, OSFD and OSFWD groups. The thickness feature provides a good sensitivity of 80.16%, specificity of 100% and positive predicative accuracy of 100%.

Some methodologies of classification used by various authors for classification of microscopic biopsy images are described in Table 2.3.

Table 2.3: Previous work reported in the literature for the classification of microscopic biopsy images

Authors (Year)	**Methods used for classification**	**Parameters used for performance measure**	**Accuracy (%)**
Metin N. Gurcan *et al., (*2009)	Support vector Machine (SVM)	F-measure, specificity.	97.00
P.W. Huang *et. al., (2009)*	Support vector Machine (SVM)	F-measure, ROC	92.88
M. Muthu Rama Krishnan *et al*., (2012)	Decision Tree (DT), Sugeno Fuzzy, Gaussian Mixture Model (GMM), K-Nearest Neighbor (K-NN), Radial Basis Probabilistic Neural Network	Accuracy, sensitivity and specificity	95.70

	(RBPNN)		
S. Di Cathaldo *et al,.* (2010)	Support vector Machine (SVM)	Sensitivity and specificity as well as F-Score	91.77
HE Lei *et. al. ,* (2012)	Adaptive artificial neural network (ANN), support vector machine (SVM), principal component analysis (PCA), multidimensional scaling (MDS), and iso-maps.	F-score, Sensitivity, and specificity.	90.00
Alsi Genctav *et al,.* (2012)	Radiating gradient vector flow (RGVF)	Weighted kappa coefficient	61.00
Yan Xu et. *al.,* (2014)	Multiple instance learning (MIL), and multiple clustered instance learning (MCIL)	Accuracy, sensitivity, ROC curve, F-measures	86.21
S. Rathore *et al,.* (2014)	Ensemble classification based on majority voting	Accuracy, ROC and sensitivity etc.	96.86
M. Muthu Rama Krishnan *et al,.* (2011)	Support Vector Machine (SVM)	Classification accuracy of normal, OSFWD and OSFD	94.06
M. Muthu Rama Krishnan *et al*., (2011)	Support Vector Machine (SVM)	Accuracy, sensitivity etc.	94.07
MR Mookiah *et al.* (2011)	Error back-propagation neural network (BPNN) and Brownian	Sensitivity and specificity,	92.31

	motion curve (BMC)		
M. Muthu Rama Krishnan *et al.* ,(2012)	Brownian motion curve (BMC), and SVM classifier	Accuracy, sensitivity etc.	88.38
M. Muthu Rama Krishnan *et al.* ,(2011)	Support vector machine (SVM) classifier to discriminate (classify) normal, OSFD and OSFWD groups	Sensitivity, specificity, and positive predicative accuracy	100

2.3 Performance Measures used for segmentation approaches

The brief descriptions of parameters used to evaluate the performance measures of segmentation approaches are as follows:

Jacard Index: If A is a ground truth and B is segmented image. Jacard similarity coefficient is represented by

$$J_{AB} = \frac{|A \cap B|}{|A \cup B|} \tag{2.8}$$

Dice coefficient: Dice coefficient is represented by

$$Dice_{AB} = \frac{2|AB|}{|A| + |B|} \tag{2.9}$$

where, A is ground truth images and B is obtained segmented images.

Tanimoto Index: Tanimoto similarity index represents the bitwise, pixel wise similarity between ground truth image segmented images.

$$T_{AB} = \frac{\sum_i (xi \wedge yi)}{\sum_i (xi \vee yi)} \tag{2.10}$$

Tanimoto index is also calculated in terms of TP, TP, FP and FN as follows:

$$T_{AB} = \frac{TP}{TP + FP + FN} \tag{2.11}$$

Accuracy: The segmentation accuracy of an algorithm is depends upon the number of correctly segmented pixels (i.e. true negative and true positive) and is calculated as follows:

$$Accuracy = \frac{TP + TN}{N} \times 100 \tag{2.12}$$

where, N is the total number of pixels presented in microscopic biopsy images.

True Positive Rate: Sensitivity is a measure of the proportion of pixels which are correctly segmented. It can be calculated using the following equation:

$$TPR = \frac{TP}{TP + FN} \tag{2.13}$$

Its value ranges between 0 and 1, where 0 and 1, respectively, mean worst and best segmentation approach. Similarly the false positive rate, true negative rate and false negative rate can be defined by using equation (2.14), (2.15), and (2.16) respectively.

False Positive Rate

$$FPR = \frac{FP}{FP + TN} \tag{2.14}$$

True Negative Rate

$$TNR = \frac{TN}{FP + TN} \tag{2.15}$$

False Negative Rate

$$FNR = \frac{FN}{TP + FN} \tag{2.16}$$

Probability Random Index (PRI): Probability Random Index is the non-parametric measure of goodness of segmentation algorithms. Random index

between test (S) and ground truth (G) is estimated by summing the number of pixel pairs with same label and number of pixel pairs having different labels in both S and G, and then dividing it by total number of pixel pairs. Given a set of ground truth segmentations G_k, the PRI is estimated using Equation (2.17) such that c_{ij} is an event that describes a pixel pair p_{ij} is ground truth probability that same or different label in the test image Stest.

$$PRI(S_{test}, G_k) = \frac{1}{\binom{N}{2}} \sum_{\forall i,j \& i<j} \left[c_{ij} p_{ij} + (1 - c_{ij})(1 - p_{ij}) \right] \tag{2.17}$$

Variance of Information (VoI): The variation of information is a measure of the distance between two clustering (partitions of elements) A clustering with a clusters is represented by a random variable A, $A = \{1.....k\}$ such that $P_i = \frac{|A_i|}{n}$, $i \in A$ and $n = \sum_i A_i$ is the variation of information between two clustering A and B. Thus VoI (A,B) is represented using Equation (2.18):

$$\mathrm{VoI}(A,B) = \mathrm{H}(\mathrm{A}) = \mathrm{H}(B) - 2\mathrm{I}(A,B) \tag{2.18}$$

where, $\mathrm{H}(\mathrm{A})$ is entropy of A and $\mathrm{I}(\mathrm{A,B})$ is mutual information between A and B. $\mathrm{VoI}(\mathrm{A,B})$, measures how much the cluster assignment for an item in clustering A reduces the uncertainty about the item's cluster in clustering B.

Global Consistency Error (GCE): The GCE is estimated as follows:

Suppose, segments s_i and g_j contain a pixel, say p_k, such that $s \epsilon S$, $g \epsilon G$ where S denotes the set of segments that are generated by the segmentation algorithm being evaluated and G denotes the set of reference segments. To begin with, a measure of local refinement error is estimated using Equation (2.19) and then it is used to compute local and global consistency errors, where *n* denotes the set difference operation and *R(x,y)* represents the set of pixels corresponding to

region x that includes pixel y. Using Equation (2.19) the global consistency error (GCE) is computed using (2.20) where n denotes the total number of pixels of the image. GCE quantify the amount of error in segmentation (0 signifies no error and 1 indicates no agreement).

$$E(s_i, g_j, p_k) = \frac{\left|R(s_i, p_k) \setminus R(g_j, p_k)\right|}{\left|R(s_i, p_k)\right|} \tag{2.19}$$

$$GCE(S,G) = \frac{1}{n}\min\left\{\sum_i E(S,G,p_i), \sum_i E(S,G,p_i)\right\} \tag{2.20}$$

2.4 Metrics used for performance evaluation of computer aided diagnostics (CAD) system for microscopic biopsy images

The performance of the automatic cancer detection system is quantitatively evaluated using well-known performance measures used for classifiers such as accuracy, sensitivity, specificity, Mathew's correlation coefficient (MCC), F-score, Kappa statistics, and receiver operating characteristics curve (ROC). Generally, a particular measure of accuracy takes into account a certain factor underlying the yielded classification results. However, one can use multiple classification measures in order to obtain more reliable comparison. Normal and malignant images, respectively, correspond to negative and positive samples. In this context, true positive (TP) and true negative (TN), respectively; represent the number of malignant and normal samples, which are correctly classified. Likewise, false positive (FP) and false negative (FN), respectively, represent the number of normal and malignant samples, which are incorrectly classified. Confusion matrix (CM) obtained by classifier is used to calculate the value of TP,TN, FP and FN. The basic definitions of these performance measures are given as follows:

Accuracy: The classification accuracy of a technique depends upon the number of correctly classified samples (i.e. true negative and true positive) and is calculated as follows:

$$Accuracy = \frac{TP + TN}{N} \times 100 \tag{2.21}$$

where, N is total number of sample.

Sensitivity: Sensitivity is a measure of the proportion of positive samples which are correctly classified. It can be calculated using the following equation:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.22}$$

Its value ranges between 0 and 1, where 0 and 1, respectively, mean worst and best classification.

Specificity: Specificity is a measure of the proportion of negative samples which are correctly classified. It can be calculated using the equation:

$$Specificity = \frac{TN}{TN + FP} \tag{2.23}$$

Its value ranges between 0 and 1, where 0 and 1, respectively, mean worst and best classification.

Matthews's correlation coefficient (MCC): MCC is a measure of the eminence of binary class classifications. It can be calculated using the following formula

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP + FN)(TP + FP)(TN + FN)(TN + FP))}} \tag{2.24}$$

Its value ranges between -1 and +1, where -1, +1 and 0, correspond to worst, best, and random prediction respectively.

F-score: F-score is a measure of the accuracy of classification. F-score is a weighted average of precision and recall, and can be calculated using the following equations (2.25- 2.27)

$$Precision = \frac{TP}{TP + FP} \tag{2.25}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.26}$$

$$F\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.27}$$

The value of F-score ranges between 0 and 1, where 0 means the worst classification, and 1 means the best classification.

Kappa statistic: Kappa statistic (k) measures the agreement between ground truth and the results of a classification algorithm .The equation for k is:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{2.28}$$

where, Pr(a) is the relative observed agreement among ground truth and the classification algorithm, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. The value of k varies between 0 and 1, where k=1 shown near to perfect agreement, and k=0 shows no agreement between ground truth and the classification algorithm.

Receiver operating characteristic (ROC) curve: The ROC curve is obtained by plotting the value of false positive rate at x-direction and true positive rate at y-direction. The area under curve (AUC) provides the measurement of effectiveness of segmentation algorithms. The range of area under curve is between, 0% to 100%. Where less that 60% AUC is poor test and greater than 90% AUC is best test for a segmentation approach.

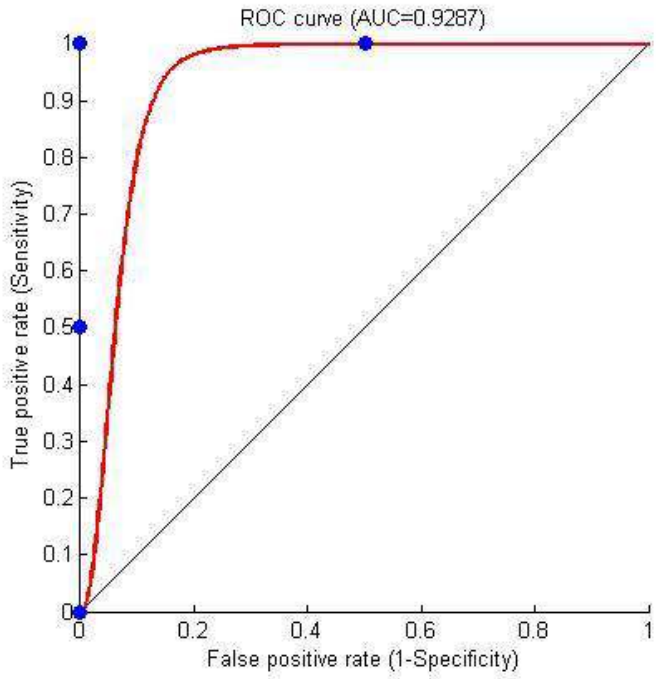

Figure 2.2: ROC of Classification approach

2.5 Data set description

For the testing and experimentation purposes, a total of 2828 histology images from the histology image dataset (histologyDS2828) and annotations are taken from a subset of images related to above database (Caicedo, J. C., Cruz, A., & Gonzalez, F. A. 2009). The image distributions based on the fundamental tissue structures in the histology data set include Connective-484, Epithelial-804, Muscular-514 and Nervous-1026 microscopic biopsy images with magnifications 2.5×, 5×, 10×, 20×, and 40×. Although the method is magnification independent, in this work the results are provided on samples digitized at 5× magnification. The features extracted from microscopic biopsy images must be biologically interpretable and clinically significant for better diagnosis of cancer. Table 2.4 provides the brief description of dataset used for identification of cancer from microscopic biopsy images.

Table 2.4: Image distribution of fundamental tissues data set of 2828 histology images

Fundamental Tissue	**Number of Images**
Connective	484
Epithelial	804
Muscular	514
Nervous	1026
Total	**2828**

2.6 Conclusion

This chapter the theoretical background of cancer detection CAD tool has been investigated to identify the problems pertaining to automated detection of cancer from microscopic biopsy images using image processing and pattern recognition tools. For the same, prominent digital image analysis and pattern recognition methods to cancer detection have been reviewed and analyzed. The problems faced by pathologist during manual analysis are also discussed. This work is presented for four significant steps of cancer detection from microscopic biopsy images using image processing and pattern recognition tools, namely image pre-processing, image segmentation, feature extraction and classification. In image preprocessing, the problems pertaining to contrast enhancement, illumination variations, and presence of artifacts/noise in the images are discussed. Segmentation is the second major step of cancer detection. For microscopic biopsy images, many segmentation methods have been used. The major problems faced in segmentation of biopsy images are the presence of artifacts, overlapped cells, and shape variations. The feature extraction also has significant steps in cancer detection from microscopic biopsy

images. Classification is used to classify the microscopic biopsy images in normal, benign and malignant. Various performance measures for segmentation and classifications are discussed. A brief description of data set description used in testing and experimentation of a CAD tools are also provided. Finally measuring metrics for performance of the overall system are also described.