Chapter 2: Literature Review

2.1 Introduction:

In this chapter the brief overview of computer aided diagnosis methods specifically discussed in context with histopathological images are presented. The motive is to discuss the recent techniques available for preprocessing, segmentation, feature extraction and classification of Haematoxylin and Eosin (HE) stained histopathological slides images of cancerous cells obtained from biopsy technique. Identification of abnormality in histopathological slides is discussed here including current challenges in this domain. With the increasing use of histopathological images for diagnosis, treatment, planning and clinical studies, it has become necessity to utilize a computer aided diagnosis (CAD) based platform for assisting pathologist experts in diagnosis of normal or cancerous samples. Reliable algorithms are essential for the extraction of morphological tissue structure and another region of interest (ROI). The goal of computer aided diagnosis is to automate the process of detection so that large no of cases can be handled in minimum time with significant accuracy.

The extensive use of CAD these days can be traced back to the appearance of digital histopathology. In recent times, CAD has become a part of routine clinical detection methods for cancer diagnosis using digitized histological images at various screening centre and hospitals and hence it has become one of the most important key research subjects in histopathological imaging and diagnostic (Chen *et al.,* 2015). Currently, manual analysis of a histology images remains the major tool for identification and diagnosis of cancerous tissue (He *et al.,* 2012). Histopathological image analysis of the cancer is currently most reliable and effective tool for cancer detection (Tabesh *et al.,* 2007). The main motive behind the application of CAD system is for identification and quantification of the diseases. Histopathology is the examination of a biopsy sample which is processed and fixed onto glass slides. The histological techniques are used for the study of structure, function, and interpretations of the tissue and cells (He *et al.,* 2012; Madabhushi *et al.,* 2009) observed the challenges in digital imaging that led to improvement in image analysis techniques resulting in improved opportunities to the pathologist for the treatment.

Pathologists visualize under microscope histopathology images for the examination of abnormalities present in the cells based on various characteristics and distributions of the cell nuclei such as size, shape, color, proportion to cytoplasm etc. In order to analyze an image for cancer diagnosis, a pathologist examines the slide manually. The pathologist analyses the spread of abnormality and the regularity of cell shape to make a diagnosis and to decide if the tissues are cancerous (Nauth, 2007).

Histopathology images consist of background, squamous epithelium and stroma cells. In the histology image, the ROI is the squamous epithelium where each cell has a nucleus and cytoplasm. Before additional processing, the background and stroma will need to be separated. To observe different structure of the tissue under a microscope, the sections of tissue sample are dyed with HE stains. The purpose of staining is to reveal the cellular structure; counter stains are used to provide color, visibility and contrast. Figure 2.1 shows that Haematoxylin stains cell nuclei blue while Eosin stains cytoplasm and connective tissues pink (Yang *et al.,* 2005).

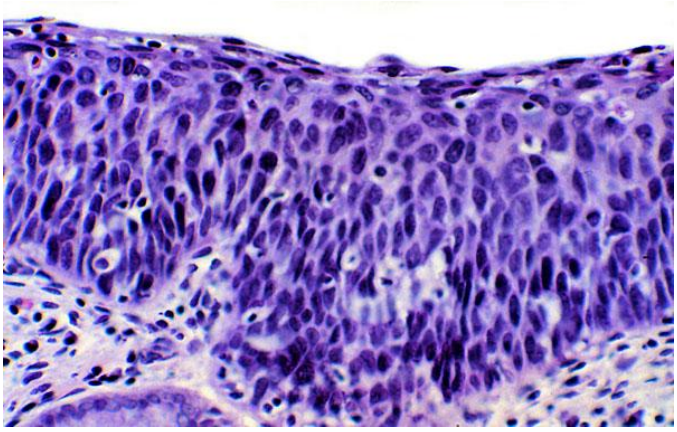

Figure 2.1: Photomicrograph of a cross section of the histopathological images of breast cancer from a biopsy sample.

In the past decades, for diagnosis of disease pathologists have examined histopathological images using manual methods. Moreover, it is found to be time consuming and tedious process. This depends on perceptions and level of expertise of pathologists (Zhang *et al.,* 2015; Gurcan *et al.,* 2009; He *et al.,* 2012).

To overcome this problem, there is the need for automation of image analysis. Therefore, computer aided diagnosis techniques are being introduced for fast, reliable and accurate diagnosis of cancer. He *et al.* (2012) proposed CAD systems are becoming crucial to improving the reliability of cancer diagnosis: a tremendous amount of research papers were conducted for automated cancer detection. This will help users and clinicians without computer training to interpret histological images and making decisions.

The computer aided diagnosis (CAD) consists of four major steps to identify the cancerous tissues. It includes preprocessing, segmentation of the histology images, feature extraction, and classification (Dougherty, 2009; Arif *et al.,* 2007; Price *et al.,* 2003; He *et al.,* 2010; He *et al.,* 2012; Gonzalez and Woods, 2004). Demir *et al.* (2005) proposed automatic diagnosis of biopsy image using image processing techniques on both tissue level and cellular level. Figure 2.2 represents the flow chart of the proposed system.

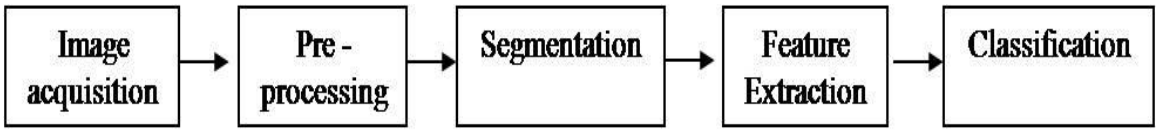

Figure 2.2: Schematic flowchart of the proposed CAD method for histopathological imaging.

In this chapter, a survey for detection and classification of cancer from histopathological images using computer aided diagnosis tools are represented. The study procedures are relevant to all imaging modalities for cancer detection from histopathology.

Doyle *et al.* (2006) proposed automated cancer diagnosis consists of four major steps: preprocessing, feature extraction, classification. Gurcan *et al.* (2009) presented a review on histopathological image analysis for the design of automated cancer detection through CAD system up to the year 2009. There is a serious need for CAD to reduce the workload on pathologists so that the experts can concentrate on the more difficult cases for diagnosis. Belsare *et al.* (2012) reviewed the different techniques used in computer aided histopathological image analysis for cancer detection and classification.

Irshad *et al.* (2014) have detailed an overview on the strategies for nucleus detection, segmentation, feature extraction, and classification on hematoxylin and eosin. They surveyed to recognize the issues relating to identification of disease from biopsy images utilizing images processing and pattern recognition tools. Arevalo *et al.* (2014) have presented a review on automatic image analysis tasks and its current trends like digital pathology in histopathology image. In a current literature review by Veta *et al.* (2014) attention has been attracted towards the primary impediment for improvement of new histopathology images based investigation strategies in scarcity of extensive, openly accessible datasets. Bhattacharjee *et al.* (2014) reported a review of computer aided diagnosis system for detection of cancer from histopathological images using image processing method.

Rajesh *et al.* (2015) presented a systematic survey on computational steps for detection of cancer from biopsy images using image processing and pattern recognition tools namely image preprocessing, image segmentation, feature extraction and classification of sample image into normal and abnormal categories e.g., benign and malignant ones.

Aswathy *et al.* (2016) discussed the applications of digital image processing techniques on histopathological images for the detection of breast cancer and its future possibilities. Different techniques used for histopathology image analysis with a focus on breast cancer classification are studied in this article. The major centre of this learning is to point out recent development in breast cancer detection and classification. Further, to provide outlook on efficiency, authenticity, and accuracy of different techniques. This paper proposes the most accurate imaging methods in the field of biopsy.

Lakshmi *et al.* (2016) summarized and analyzed various procedures on detection of breast cancer. Their work suggested that the CAD software works to elevate the spots that may indicate cancer so the radiologist can then look more closely at the mammogram. The imaging modalities and the imaging mining concepts are also discussed.

Chen *et al.* (2017) reviewed the haematoxylin and eosin histopathological images of breast cancer using CAD system. They reviewed the systematic procedures of image analysis for breast cancer prognosis, including image acquisition, image preprocessing, image detection with segmentation, and feature extractions. They also evaluated the prognostic value of image features and image feature based prognostic models.

Jothi *et al.* (2017) explored the materials and methods that are being used or previously has been used for CAD to distinguish cancer from histopathological images. The study was done to find the most prominent technique at different step using pie charts of histopathological image analysis.

Gopalakrishnan *et al.* (2017) presented a detailed review of the automated analysis of histopathology images for cancer diagnosis. Their work discusses the brief study of algorithms used for preprocessing, segmentation and feature extraction and also their classification. The work was done in identifying various issues that needed to be addressed and solved for building the robust automated system. The steps to be followed in developing an automated system for the diagnosis or detection of malignant cancer from histopathology images were studied briefly. Results of the study suggest that ensemble based learning system give improved recognition rate and boundary based nuclear feature helps for better classification rate.

Jalalian *et al.* (2017) summarized the approaches that were applied to design different stages of CAD system. Further, the advantages and disadvantages of different segmentation, feature extraction, and classification techniques are discussed here. The review also discusses the impact of imbalanced datasets in classification outcomes and appropriate methods to solve these issues. As well as, performance evaluation metrics for various stages of breast cancer detection in CAD systems are also reviewed. Their study suggested that the region based segmentation and clustering-based algorithms are being widely used to develop CAD systems for breast cancer detection. Further, the paper proposes the techniques for extracting suitable features for the detection of normal and abnormal lesions in breast depending on the nature of mass and imaging modalities.

The present thesis primarily focuses over imaging slides as acquired from the histology images, particularly biopsy sample images, as they are considered to be the gold standard in cancer identification and diagnosis. Furthermore, the acquired histology images are processed utilizing the techniques as represented in the next sections. Figure 2.3 represent major step in the detection of cancer cells by using computer aided diagnosis of histopathological images and elaborate the techniques utilized for diagnosis of cancer diseases.

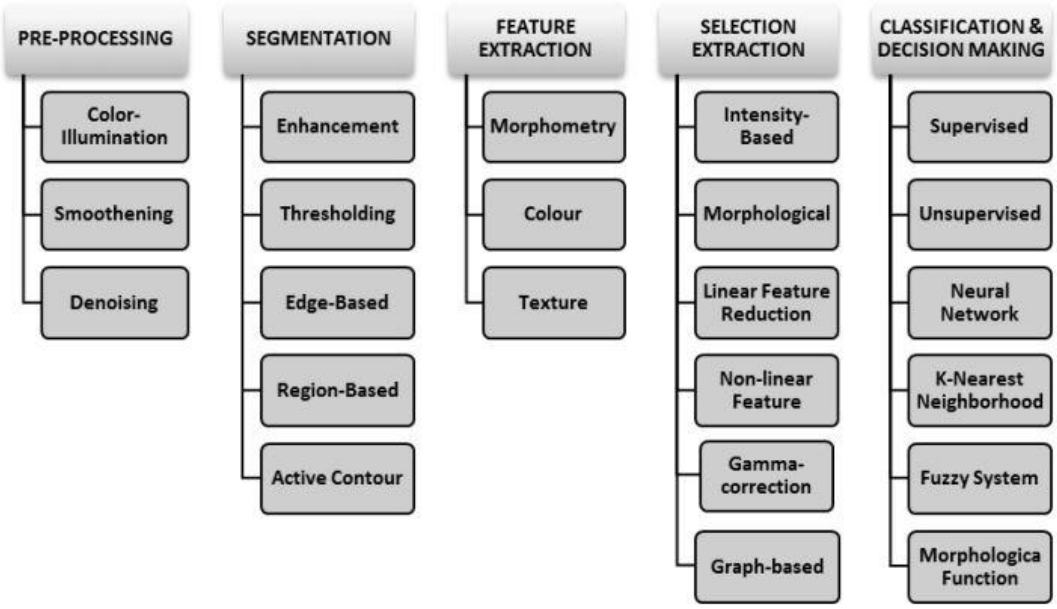

Figure 2.3: Schematic diagram of methods used in CAD method for histopathological imaging.

2.2 Preprocessing:

Histopathology images are acquired from microscopes, henceforth, it may consist some deficient like uneven staining, low contrast, dust particles, air bubbles, and tissue folding. Pathologists face difficulty in cell image detection because of overlapping, blurred, low contrast as well as weak boundary detection caused by uneven dyeing. Preprocessing step decrease the background staining and improve the image quality (Dougherty, 2009; Stoitsisa *et al.*, 2006; Li and Najarian, 2008).

Rahmadwati *et al.* (2010) proposed that 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 (object of interest) and background.

Preprocessing includes several methods such as unwanted noise removal, image sharpening, smoothening, deblurring and edge enhancement etc. Large number of techniques has been applied on the enhancement of gray level image (Takashi *et al.,* 2009). Many medical images consist of a background and region of interest (ROI) and pre-processing technique usually separates these images (Plissiti *et al.,* 2011; Luck *et al.,* 2003; Zhang *et al.,* 2000). Subsequently, a thresholding process is employed for image enhancement (Plissiti *et al.,* 2011). Images undergo noise filtrations for better detection of cell and nuclei present in the cancerous cells for image classification purposes (Plissiti *et al.,* 2011; Praveen and Vijayan, 2006). Numerous methods, such as thresholding and adaptive filtering, have been reported to eliminate and filter noise (Gonzales and Woods, 2007; Praveen and Vijayan, 2006; Plissiti *et al.,* 2011; Zhang and Liu, 2004). Filtering methods, for example, Low pass filter (LPF) and median filter, Gabor Filter, Gaussian Filter and Prewitt edge finding filter usually filters the noises in an images (Rahmadwati *et al.,* 2010; Gonzales *et al.,* 2007; Li *et al.,* 2008; Dermir *et al.,* 2005; Luck *et al.,* 2003).

Contrast enhancement is an important factor for image enhancement. In this method, contrast of an image is improved to make the image better for human vision. The term contrast, is defined in digital images, is the separation of dark and bright areas present in the image (Gonzales and Woods, 1992). Contrast enhancement methods include gamma correction, single-scale retinex, histogram equalization, Contrast limited Adaptive Histogram Equalization (CLAHE). Tay *et al.* (2012) after detecting a region of interest (ROI), and then a gamma correction was applied to the red channel of the ROI. Gamma correction can change the brightness and the ratio of red, green, and blue of an image; it highlighted the cell nuclei from the background. Jobson *et al.* (1997) have reported a retinex theory that also leads to contrast enhancement of an image.

CLAHE algorithm improves the image contrast by improving the local contrast present in an image and by enhancing the weak boundary edges in each pixel of an image through limited amplification (Zuiderveld *et al.,* 1994). Haematoxylin and Eosin staining images have been enhanced by using contrast enhancement by using Contrast Limited Adaptive Histogram Equalization (CLAHE) techniques. Stained images in color hence enhancement done on individual color spaces by sampling, subsampling and color space conversions (Prasoon *et al.,* 2013; Gurcan *et al.,* 2009) proposed artifacts found in digital pathology. The basic task in preprocessing is nose reduction and contrast enhancement of the region of interest. HE staining images have been enhanced by using contrast enhancement by using CLAHE techniques. Singh *et al.* (2012) applied CLAHE methods in biopsy image of cancer cells detection and classification.

In the last three decades of literature review includes various aspects of non-linear dynamics involving Stochastic Resonance (SR) phenomenon. Previous proposed techniques for noise removal or contrast enhancement has not produced desired contrast levels and distorts the image contents. Ye *et al.* (2004) had applied SR phenomenon in the context of image quality enhancement of low-contrast sonar images. Author have reported that SR based enhancement technique is supportive to enhance low contrast images affected from noise. Peng *et al.* (2007) studied a novel preprocessing approach to improving the low-contrast medical images using SR.

Gammaitoni *et al.* (1998) and McNamara *et al.* (1989) extensively studied the concept and comprehensively reviewed on Dynamic Stochastic Resonance (DSR). Adding some suitable noise to the input image enhancement can improve. Using DSR an analogy of a low contrast image to a bi-stable double well dynamic system has reported and the poor state of pixel gray values has made to transit into enhanced state. However, current studies have persuasively shown that in nonlinear systems, noise induce more ordered regimes that basis the amplification of weak signals, and increase the signal to noise ratio (SNR).

Also, the technique reported in Peng *et al.* (2007) used non DSR to improve the performance of Adaptive Histogram Equalisation (AHE) by using SR. Rallabandi *et al.* (2008) successfully enhanced the ultrasound images by using wavelet transform as the input to DSR and various lesions in MRI images were enhanced in which they applied Fourier transform as the input to DSR (Rallabandi *et al.*, 2010). Xiao *et al.* (2007), Peng (2007), Ye (2004), Piana *et al.* (2000), and Simonotto *et al.* (1997) have been reported the application of SR for grayscale image or edge enhancement processes. Rallabandi *et al.* (2010) proposed Quartic bi-stable model of SR with Fourier transform coefficient of images and found valuable for diagnosing of brain lesions in Magnetic Resonance Imaging (MRI).

2.3 Segmentation:

Segmentation is the process where an image is divided into the different regions on some similarity bases or divided into its constituent objects, or parts, and background (He *et al.*, 2012). Segmentation is the process of dividing an image into regions with similar properties such as gray level, color, texture, brightness, and contrast (Gonzalez and Woods, 2004; Pratt, 2001; Pal and Pal, 1993).

Although a number of algorithms have been proposed in the field of histopathology image segmentation, histopathology image segmentation is still a complex and challenging problem. In recent years, many of the segmentation methods have been registered for the cell, nuclei and cytoplasm detection on histopathology images using threshold, region based and clustering based algorithms. However, the extraction of a region of interest (ROI) like a cell, nuclei and cytoplasm decide to select appropriate segmentation approaches (Gurcan *et al.*, 2009; Gonzalez and Woods, 2004; Sezgin and Sankur, 2003; Loukas and Linney, 2004).

Thus, the broad classifications of techniques available for segmentation of an image, classified into three classes as follows: (a) Pixel based segmentation (histogram), (b) Edge based methods, (c) Region based methods. (Figure 2.4).

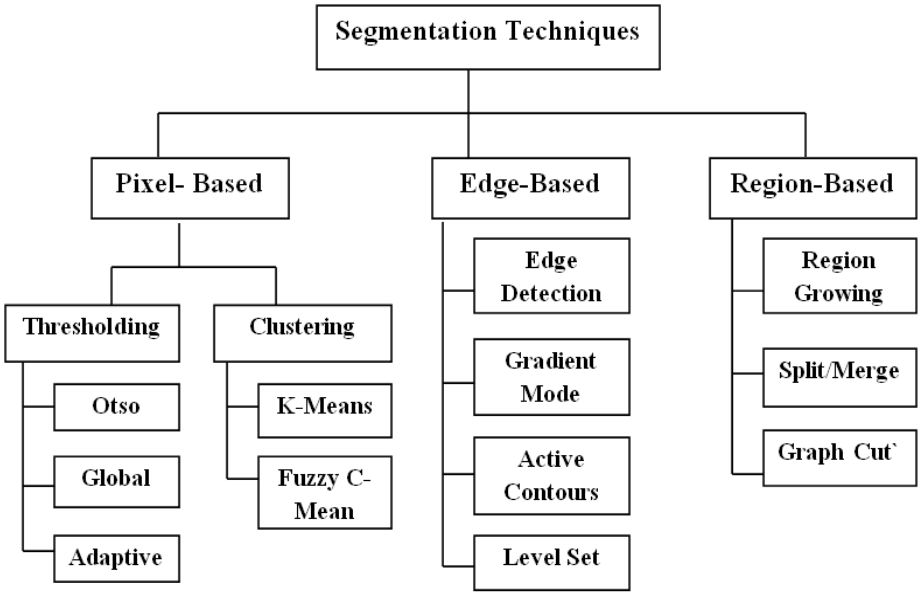

Figure 2.4: Classification of segmentation techniques.

The detail of each methods as follow:

2.3.1 Pixel based segmentation:

Segmentation of an image based on thresholding of gray level thresholding and histogram features. This is appropriate for image having object or region of uniform brightness located adjacent to a background of dissimilar gray level. A threshold applied to segment the object and the background. For nuclear segmentation pixel based method are the used. They are based on the information of the pixel value of the gray level, color, and texture etc. Rather than pixel based methods, the thresholding techniques use 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 (He *et al.,* 2012; Kim, *et al.,* 2007). Naik *et al.* (2008) divided pixel values into low level and high level information between object pixels and the background pixels discrimination to segment cancer cells. Thiran *et al.* (1996) authors reported on the pixel of the normal and cancerous nucleus in histopathological images. Guillaud *et al.* (2004) has examined segmentation of cervical histology images. The tissue image was segmented based on intensity information, edge magnitude information and edge connectivity information.

Researchers have segmented the histology images using a threshold algorithm (Keenan *et al.*, 2000; Loukas *et al.*, 2004; Gonzales *et al.*, 2008). In thresholding, shape-based separate the object from the background, a value based on image intensity is used. Keenan *et al.* (2000) suggested an incremental thresholding method for segmentation of the nuclei. He *et al.* (2010) presented an algorithm using Gaussian mixture modelling (GMM) for segmentation of haematoxylin and eosin stained cervical histopathology image. This method extracted the tissue constituents such as nuclei, cytoplasm, blood cells and stroma from the background.

Wilkinson *et al.* (2003) proposed segmentation method using robust automatic threshold selection (RATS) for microbes image analysis, in that they have reported that RATS is suitable for thresholding a noisy image with the variable background. Bredfeldt *et al.* (2014) have demonstrated a protocol that allowed consistent scoring throughout large patient cohorts in two steps; the first step involves the use of Trainable Weka Segmentation (TWS) Image J plugin for finding epithelial cell nuclei and other involves the application of a cascaded matched filter, threshold operation to identify clusters and boundaries. Liu *et al.* (2007) introduced the microscopic image analysis with the help of adaptive threshold segmentation. For nuclei detection, most of the researchers used Fuzzy based technique Pham *et al.* (1999) best suited for nuclei segmentation.

2.3.2 Edge based segmentation:

Edge based segmentation is the common method based on detection of edges hence, boundaries which divide different regions. Edge detection method is based on the discontinuities in gray level, colour etc., and these edges correspond to boundaries between objects. Thus edges based method divide an object on the basis of boundaries, i.e. edges are often used in image analysis to construct image boundaries (Brejl and Sonka, 2000).

2.3.3 Region based segmentation:

Region based method are based on the principle of homogeneity, and the pixel having similar properties are clustered together to form a homogenous region (Gonzalez and Woods, 2004). The criteria for homogeneity are most of the time texture, intensity (gray level of pixels), and color (Sonka *et al.*, 2000; Wang *et al.*, 2008; Gonzalez and Woods, 2004). Region based segmentation methods include (1) region growing (Hebert *et al.,* 1996) and (2) clustering (He *et al.,* 2010; Gonzalez and Woods, 2008). Region segmentation based on intensity similarity is used in medical image segmentation such as region growing (Dawei *et al.,* 2007; Zhang *et al.,* 2009). The principal of region growing relies on tracking neighbouring pixels based on a similarity and discontinuity measurement.

Tosun *et al.* (2008) proposed a homogeneity measure based on the distribution of the objects that are defined to represent tissue components. Using this measure, they demonstrated a new object oriented segmentation algorithm and therefore, implemented the object oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection.

Al-Kofahi *et al.* (2009) authors presented two automated methods for the segmentation of histology tissue images to overcome the limitations of the manual approach and also of the existing computerized techniques. The research article represents, first independent technique depending upon unsupervised colour clustering that automatically detects primary goal based cancerous zones in the biopsy samples and disregards the stroma. The second technique depending upon colours separation and morphological examination uses automated segmentation of the cancerous nuclear membrane cells. The experimental results of the study demonstrated the accuracy of the techniques compared to manual segmentations; it also proposes that these techniques are more effective in histology images than popular approaches based on supervised learning or active contours. Their work suggested that presented procedure can be used 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.

Huang and Lai (2010) represented a confined and precise technique for segmenting cell nucleus utilizing amalgamation of information. The central images were extracted utilizing the graph cut based algorithms. Afterwards, through merging of multiple Laplacian of Gaussian (LoG) approaches the nuclear seed points were screened. These points are utilized for executing introductory segmentations and once more polished utilizing the second graph-cut approaches. This algorithm utilizes 25 image biopsies along with 7400 nuclei and segmentation errors were examined here. In general, as reported the accuracy level of the algorithm has been near to 86%.

Dundar *et al.* (2011) utilizes edge detection as feature detection for characterizing the cell size under segmentation approaches. They utilized the significant lab expert opinions for detection of Oral Submucous Fibrosis (OSF).

Di Cataldo *et al.* (2012) authors worked on an innovative synergistic boundary and region based active contour model that includes shape priors in a level set formulation with automated initialization based on the watershed. Research work demonstrated an application of these synergistic active contour models using multiple levels sets to segment nuclear and glandular structures on digitized histopathology images of breast and prostate biopsy specimens. In comparison to previous associated approaches, proposed model is able to resolve object overlap and separate occluded boundaries of multiple objects at the same time. The results of qualitative and quantitative evaluation on 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.

Genc-tav and Aksoy (2012) authors proposed the MCIL method which simultaneously performs image level classification (cancer vs. non-cancer image), medical image segmentation (cancer vs. non-cancer tissue), and patch-level clustering (different classes). It implants the clustering notion into the multiple instances learning (MIL) setting and derives a principled solution to performing these three tasks in an integrated framework.

Research work also includes a contextual constraint as a prior for MCIL, which further reduces the ambiguity in MIL. Experimental results of the study displayed on histopathology colon cancer images and cytology images verified the great advantage of MCIL over the competing methods and the accuracy was 98.92%.

Mouelhia *et al.* (2013) authors presented an algorithm for semi automatic segmentation of the nuclei under the adequate control of the expert user. This algorithm can work automatically 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 the web-based architecture it uses. The algorithm localise the cell nuclei applying a voting pattern and preceding information before identification of exact shape of the nuclei through elastic segmentation algorithm. The noises are filtered using mean-shift and median shift. The edges are extracted utilizing the canny edge based algorithms. The edges near to back ground are filtered as cell nuclei are enclosed within cytoplasm and their possessing rough elliptical shapes.

Rathore *et al.* (2014) authors presented an automatic system to achieve both segmentation of touching nuclei in order to get the total number of cancer nuclei in each class. In the study, a modified geometric active contour model is used for multiple contour detection of positive and negative nuclear staining in the microscopic image. They also proposed a touching nuclei method based on a watershed algorithm and concave vertex graph to perform accurate quantification of the different stains. The results exhibited the superiority of the suggested methods when compared with other current 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.

The table 2.1 lists a brief introduction about histopathological image segmentation approaches and data set used for segmentation.

Table 2.1: Previous work reported for the segmentation of histopathological images.

S. No.	Authors (year)	Methods used for segmentation	Object for segmentation	Data set used
1	Gurcan *et al.*, 2009	Shaped-based approaches, active counter, fuzzy c - means and watershed shed method	Cell Nuclei, cell detection rate achieved about 90-90%	20 biopsy of size 512×512×512 of 4 billion pixels
2	Huang *et al.*, 2009	Marker control watershed transform	Nucleus and cytoplasm	1000×1000,4000×3000 and 275×275 HCC biopsy images
3	Ghosh *et al.*, 2010	Fuzzy divergence and modified thresholding	Segmentation of cancerous and non-cancerous tissue	Biopsy images
4	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
5	Tosun *et al.*, 2011	Graph Run length matrices	99% accuracy For segmentation	Biopsy images
6	Basavanhally *et al.*, 2011	Active Contour model based on Color Gradient with Hierarchical Normalized cut	89% segmentation accuracy	Histopathology images
7	Dundar *et al.*, 2011	Gaussian Mixture Model based spatial information and Expectation Maximization (EM) Algorithm	Overall accuracy of 87.9%	Oral Submucous Fibrosis (OSF).
8	He *et al.*, 2012	Thresholding, active contours and Markov random fields (MRF)	Nuclei, cytoplasm, stroma	Digitized histology images

9	Cen *et al.*, 2012	Thresholding, watershed, and active contours	Nuclei, area, radix index with 94% accuracy	Microscopic biopsy images
10	Mouelhi *et al.*, 2013	Geometric active contour, watershed	Nuclei touching and nuclei cytoplasm segmentation with accuracy 96.5%	2048 × 1360 JPEG files with 24 bit/channel
11	Kowal *et al.* 2013	Adaptive thresholding, k-means, fuzzy c means (FCM) and k-nearest neighbor (KNN)	Nuclei and cytoplasm with 96% segmentation accuracy	704 × 578, BMP image files with 8 bit/channel RGB.
12	Xu, *et al.*, 2014	Multiple clustered instance learning (MCIL)	Segmentation of cancerous and non-cancerous tissue with 96% accuracy	Sample biopsy images
13	Xu *et al.*, 2014	Conditional random field (CRF) based approach	Nuclei segmentation with 98% accuracy	60 RGB biopsy images
14	Veta *et al.*,2014	Marker-controlled watershed algorithm and Fast radial symmetry transform	81.2% accuracy	Breast cancer histopathology images
15	Jain *et al.*, 2014	Active Contour Model (Both region based and boundary based) with General Classifier Neural Network (GCNN)	83.47% accuracy	Histopathology images

2.4 Feature extraction:

Feature extraction is a crucial and challenging aspect in the CAD and many image processing and computer vision applications, such as image retrieval (Antani, 2002; Veltkamp, 2001), registration and matching (Zitová *et al.*, 2003), and pattern recognition (Bishop, 2006). Features extraction is needed to quantify the cellular changes in the tissue. The feature extraction is applied for the cell level or the tissue level to determine the properties of image deformity or to assign the histological image to its pattern (Belsare *et al.*, 2012). The objective of feature extraction is to reveal all possible features from input data that are expected to be efficient in diagnosis (Rodenacker *et al.*, 2003; Zheng *et. al* (2017).

In recent years, many machine learning methods have been introduced to extract features from histopathological images.

2.4.1 Morphological features:

Morphological features can be used to differentiate between benign and malignant cells by quantifying the properties of size and shape followed by statistical measurement. Morphology features involve parameters concerning shape and size of the ROI. In histology images, the associated and extracted features are used to analyse the shape and size of a cell. Morphology features may be obtained and then measured by using a bounding box (which is useful to derive position and orientation features), geometric measurement (including features such as cell area, perimeter and radius), contour parameters (such as cell curvature, bending energy, a convex hull, and elliptic deviation) and invariant moments (Guillaud *et al.*, 2004). Morphological features can be extended to include biological features such as the nuclei to cytoplasm ratio (N/C). Anderson *et al.* (1997) worked on morphological features of glandular tissue components for separation of benign and malignant tumour in a breast tissue. Guillaud *et al.* (2004) computed morphological features for quantitative analysis of CIN cervical histological images.

Mulrane *et al.* (2008) has shown the quantification of the morphological features and classify their structure in a histopathological image leading to identify of a cell into a particular class for the reason of diagnosis.

Gurcan *et al.* (2009) documented CAD based approach for automatic cell segmentation and nucleus-to-cytoplasm quotient investigation. The investigational findings revealed that the technique produces particular segmentation with high effectiveness and steady accurateness. Furthermore, the examined NC quotient components are very near to the findings of manual cell segmentation. This fact reveals that the algorithm proves to be significant and competitive from the clinical point of view. Henceforth, the CAD based algorithms has proved to be beneficial for biomedical image based investigations.

Basavanhally *et al.* (2010) worked on the random variable denoting features the square root of the area and standard deviation that helps to measure the intensity for each region. Rahmadwati *et al.* (2010) use the morphological method for quantifying the characteristics of normal and abnormal nuclei in cervical histology images.

Monaco *et al.* (2010) gland area was used as discriminative criteria to categorize between benign or malignant. As well as (Naik *et al.*, 2007) and some of mentioned works, most of the morphological measures estimation are done based on a previous segmentation, and therefore its performance depends on the precision of such segmentation.

Belsare *et al.* (2012) studied on the tissue structure and cell distribution in a tissue, they mentioned irregularities of the size and shapes of cells to conclude the level of malignancy in histopathological images. Examination of the cells based on their morphological difference is useful to study the discrimination of benign and malignant cells.

2.4.2 Intensity features:

Intensity based features provides information of the gray level or color of pixels located in the ROI. Intensity based features are apply for describing pixel level characteristics of images. Gray level or color histograms of intensity values and densitometric features are employed. Properties of the features such as mean, standard deviation, kurtosis, and skewness are computed to find first order statistical information on the texture of tissues. This feature extraction approach uses different color spaces.

Weyn *et al.* (1998) employed optical densities of nuclear tissue components in the quantification of breast tissue images. The same group Weyn (1999) used features obtained from densitometric properties of nuclear tissue components together with morphological and structural image features in the diagnosis and prognosis of malignant mesothelioma. Wiltgen *et al.* (2003) employed co occurrence matrices to describe texture characteristics of histopathological images of skin tissues. Demir *et al.* (2005) the intensity based method is employed to calculate the intensity value of pixels to define the features in a histopathological image.

Petushi *et al.* (2006) obtained the intensity of the pixels that are registered and calculate the mean of the neighboring pixels. Samsi *et al.* (2012) used the Hue channel from the HSV (Hue-Saturation-Value) color space conversion of the original image, while Oszdemir *et al.* (2012) worked on white/pink/purple color dimension.

2.4.3 Texture features:

The texture is defined as something consisting of mutually related elements (Sonka *et al.,* 1999). A texture may be fine, coarse, smooth, or grained, depending upon its tone and structure, where the tone is based on pixel intensity properties, the structure is the spatial relationship between pixels (Haralick, 1979). Further, the spatial arrangement of texture primitives or texture element is termed as textone (Julsez, 1981). It is arranged in more or less periodic manner, where texture primitive is a group of pixels representing the simplest or basic sub pattern.

In contrast, Gonzales and wood summarise three approaches to textures analysis: statistical, Syntactic and spectral. It is not a significant differentiation. The spatial distribution of the pixel's gray level is measured by a statistical approach which applies the co-occurrence matrix (Haralick, 1979). GLCM is one of the methods in the co-occurrence matrix that applies the second order statistical information of an image (Esgiar, 1998; Hamilton *et al.,* 1997; Wang, 2007). It provides a large number of features that can be extracted from GLCM such as energy, entropy and cluster information. The variable distance and angle of the co occurrence to be considered can be chosen and the decision will influence the optimal texture characterisation. In effect it produces a large number of total features Run Length Matrices (RLM) defines the co occurrence with an unbroken pixel run that has the gray level value in a current direction. It is similar with GLCM, the image needs to be quantified, and the number of parameters that are chosen affects the texture characterisation and the classification results.

Features extracted from these colour co-occurrence matrices are also used for representing the image texture Co-occurrence matrices are also proposed to be used in quantifying histopathological images of different types of tissues (Weyn, 1999; Doyle *et al.,* 2007). 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). Some significant works had been contributed by various researchers for microscopic biopsy image analysis that contains a variety of features to derive clinically significant information.

Hamilton *et al.* (1987) reported semiautomatic image analysis to carry out a morphometrical assessment of 15 nuclear and cellular variables in normal and malignant colorectal epithelium. In another study by Hamilton *et al.* (1987) texture analysis was employed to develop criteria for the automatic identification of focal areas of colorectal dysplasia from a background of histological normal tissue. Esgiar *et al.* (1998) worked on co occurrence features for automated categorization of normal and cancerous colonic mucosa.

Demir *et al.* (2005) reported the use augmented cell-graphs, that are undirected, weighted, and complete probabilistic graphs without self-loops, for automated diagnosis of cancer, where all possible edges between each pair of nodes are included in the graph, preventing the loss of any existing spatial information. In these graphs, edge weights are defined as the Euclidean distances between their end nodes. Furthermore, Gunduz-Demir (2007) investigated the relation of different phases of cell-graphs with the malignancy of cancer using graph evolution technique.

Doyel *et al.* (2007) suggested analysis of breast cancer histopathology image by using textural and nuclear architectural features. In (Doyal *et al.,* 2008) calculated 6 Haralick features were calculated to combine them with topological and color features to grade breast cancer.

Micheletti *et al.* (2007) categorized tumour cells depending upon statistical shape analysis algorithm. They applied the size function theory and attached some statistical techniques of separate analysis, to execute automatic categorisation of random shape families. The approach is utilized for categorising normal and cancerous cell nuclei, elaborated through their sectional characteristics. The outcome of this technique is evaluated with respect to other approaches of shape analysis that were previously applied over the similar data, depicting acceptable advancements.

Krishnan *et al.* (2010) proposed a wavelet-based texture classification for oral histopathological sections. The conventional method involves in stain intensity, inter and intra-observer variations leading to higher misclassification error. The proposed method, involved feature extraction using wavelet transform, feature selection using Kullback – Leibler (KL).

Chaddad *et al.* (2011) extracted Haralick's texture features and morphological parameters from segmented multispectral texture bio images for classification of colon cancer cells. They apply the Probabilistic Neural Network (PNN), the activation function for the function that measures the distance of unknown variable to all known class variables.

Eid and Landini (2011) authors presented a number of inventive techniques to evaluate 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 study was performed 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.

Krishnan *et al.* (2011) authors presented a quantitative microscopic approach for discriminating inflammatory and fibroblast cells of oral sub mucous fibrosis (OSF) from normal oral mucosa (NOM) in order to shape features of the sub-epithelial connective tissue (SECT) cells. Segmentation and classification of sub-epithelial connective tissue (SECT) cells were used except for endothelial cells in the oral mucosa of normal and OSF conditions had been reported. The shape features used were statistically significant using Mann-Whitney U test, which enhanced the statistical learning potential and classification accuracy of the classifier.

Krishnan *et al.* (2012) authors introduced a quantitative microscopic approach for discriminating oral sub mucous fibrosis (OSF) from normal oral mucosa (NOM) with respect to morphological and textural properties of the basal cell nuclei. The experimental results revealed that 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 oncopathologists. 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.

Tam *et al.* (2016) proposed a fully automated stain normalization method to reduce batch effects and thus helped research in digital pathology applications. Their suggested method, intensity centering and histogram equalization (ICHE), normalizes a diverse set of pathology images by fist scaling the centroids of the intensity histograms to a common point. Afterwards applied a modified version of contrast limited adaptive histogram equalization. The proposed method was evaluated based on image intensity values, quantitative features, and the effect on downstream applications, such as a computer aided diagnosis. Three methods from the literature were reimplemented and evaluated using the same criteria for comparison. The results of the study suggested that ICHE not only improved performance compared with un-normalized images but in most cases showed improvement compared with previous methods for correcting batch effects in the literature.

Vaibhav *et al.* (2016) studied the histopathological biopsy image classification based on convolutional neural network. To classify the histopathological images either as benign or malignant tumors he employed Convolutional Neural Network (CNN) classifier. For this study, lung tissues were taken to demonstrate the computerized tumor detection which replaces the traditional method of observing the histopathological images under the microscope. The methodology used in this study consisted of two phases - training and testing phase, which included steps: Preprocessing, feature extraction of biopsy image using GLCM and morphological feature, PCA reduction and finally implementation of CNN classifier.

Zeng *et al.* **(**2017) worked on a patterns and distribution information of nuclei for histopathological images feature extraction framework. In which nucleus is first detected from images and trained developed by a convolutional neural network with three hierarchy structures were extracted step-by-step from the patch level, block level and image level structure of the integrated network. The study was done for 450 histopathological images.

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 histopathological images.

S. No.	Authors (year)	Methods used for feature extraction and selection	Feature extracted	Data set used and performance measure
1	Madhbhushi *et al.,* 2007	Graph-based features extraction	Shape, size, centre of mass, texture feature, and spatial related features	Number of nodes, the number of edges, sensitivity and accuracy etc.
2	Gurcan *et al.,* 2009	Graph-based features extraction	Size, shape, statistical and texture features	Number of cells, number of triangles, and N/C
3	Landini *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
4	Rahmadwati *et al.,* 2010	Comparing feature of nuclei, shape factor and roundness of nuclei	Classification of cervical cancer using Histology images	TP, TN, FP and FN ROC and AUC
5	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
6	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

7	He *et al.*, 2012	Morphometry, topological intensity, color texture	Area, size,boundary, shape, moments, Harallick's and Gabor texture features, Markovian, run length texture, wavelet density features etc.	TP, TN, FP and FN
8	Belsare *et al.*, 2012	Texture, graph, morphological, and Voronoi diagram features.	Smoothness, coarseness, regularity, correlation, contrast, the number of nuclei, shape size and roundness	Normal, abnormal, and grade of cancer.
9	Krishnan *et al.*, 2012	Texture morphological and intensity-based features were extracted	Eighteen morphological, 4 gray-level co-occurrence matrix (GLCM) based texture features and one intensity feature	OSF is selected from NOM
10	Onder *et al.*, 2013	Texture feature,	Variance, kurtosis, mean value of pixels, entropy, energy, contrast, co-relation etc.	Ground truth atlas data, ROC etc.
11	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 an accuracy of 98.85%.
12	Kumar *et al.* 2015	Texture, shape and morphology , HOG, wavelet color, Tamura's feature,and LTE	Size, shape, intensity and texture	2828 histology images accuracy = 92.19 sensitivity = 94.01 specificity = 81.99

2.5 Classification:

The features are extracted from an image as input to the classification stage. The classifier is constructed based on the inherent characteristics of an image, to discriminate the characteristics of nuclei and stages in breast histology images. The classifier is constructed according to the feature extraction. A classifier refers to an algorithm, sometimes a mathematical function that implements a classification (Chen *et al.,* 2012; Veta *et al.,* 2014).

The classifier uses the set of features and points out the difference between normal and diseased cell (Padmapriya *et al.,* 2016; Shenbagarajan *et al.,* 2016; Wang *et al.,* 2016; *Veta et al.*, 2014; Dheeba *et al.*, 2014; Aziz malaker *et al.,* 2015). Supervised algorithms include Ada Boost (Malladi *et al.,* 1995) Support Vector Machine (SVM) (Wang *et al.,* 2007; Huang *et al.,* 2008; Plissiti *et al.,* 2011),decision tree (Yu *et al.,* 2010; Vasantha *et al.,* 2010) Rajendran and Madheswaran (2010), and Markov Random Field (Luck *et al.,* 2003).

Unsupervised algorithms include k-means clustering (Wu *et al.,* 2007; Ilea and Whelan (2006) and Fuzzy c-means (FCM) (Hafiane *et al.,* 2008; Rezaee *et al.,* 1998). Many classifiers have been used for the classification purposes, some commonly used classification methods are artificial neural networks (ANN), Bayesian classification, K-nearest neighbour classifiers, support vector machine (SVM) (Sinha, 2003). SVM is widely used for classification (Wang, 2008; Wang *et al.,* 2007; Huang *et al.,* 2007).

The goal of this method is to find a set of optimal hyperplane boundaries that are classified into multiple clusters. SVM is applied to classify different types of Pap smear cells based on a morphometric characteristic of the cell (Huang *et al.,* 2008). SVM is applied by Yinhai for classifying tissue regions in the cervical histology image through the grid search method and the type of cell based on the morphometric characteristic of the cervical cell (Wang *et al.,* 2007). However, the SVM method leaks in computational cost in the selection of the parameter to obtain optimal performance. Immunohistochemically stained colon mucosa allowed for an accuracy of 92% in classification of benign and malignant images by Esgiar *et al.* (1998).

Sinha and Ramakrishnan (2003) extracted some features of microscopic biopsy images which included eccentricity, area ratio, compactness, average values of color components, energy entropy, correlation, the area of cells and nucleus. The classification accuracy achieved by Bayesian,-nearest neighbour, neural networks, and support vector machine was 82.3%, 70.60%, and 94.1%, respectively.

Landwehr *et al.* (2005) have developed algorithms for accurate and compact classifiers by evaluating the performance of logistic modal tree (LMT) on 36 datasets collected from the UCI repository. Landini *et al.* (2006) presented a method for morphologic characterization of cell neighbourhoods in a neoplastic and preneoplastic tissue of microscopic biopsy images. Work done used watershed transforms to compute the cell and nuclei area and other parameters. The distance measure of the neighbourhood value was used for the calculation then neighbourhood complexity with reference to the v-cells. The best classification which was obtained by NN classifier was 83% for dysplastic and neoplastic classes and 58% of correct classification.

Kim *et al.* (2006) used normalized mutual information index for feature selection and supported vector machines (SVM), Cox-proportional hazard regression model, and artificial neural network classifiers for classification in a sample size of 679 patients (the recurrence prevalence of 28.6%). The features used in their prognosis system were local invasion of tumor, number of tumors, number of metastatic lymph nodes, the histological grade, tumor size, estrogen receptor, and lymphovascular invasion and reached the sensitivity, specificity and area under the curve of 89%, 73% and 0.85, respectively for the best classifier (SVM). Although the statistical power of their system was acceptable (Power = 89% N 80%), the Type I error was beyond the acceptable range ($\alpha = 0.17$ N 0.05).

Gupta *et al.* (2008) proposed an innovative method of feature selection using mean -shift and recursive feature elimination techniques to increase discrimination ability of the feature vectors. The performance of the algorithm was evaluated on an in-vivo recorded LIF data set consisting of spectra from normal, malignant and pre-malignant patients. The proposed method had sensitivity of above 95% and specificity of above 99% towards malignancy.

Caicedo, (2009) authors have extracted features of 1502 histology images with 18 different concepts. The classification strategy used is based on binary classifiers following the one-against-all rule. 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 were calculated on the test dataset and averaged over all 18 classes.

Huang and Lai (2010) presented a methodology for segmentation and classification techniques for histology images based on texture features and a maximum classification accuracy of 92.8% was obtained by using SVM.

Di Cataldo *et al.* (2010) authors suggested a new learning method, multiple clustered instance learning (MCIL), for histopathology image segmentation. An integrated framework was developed to classify histopathology images as having cancerous regions or not, segment cancer tissues from a cancer image, and cluster them into different types. The system suggested automatically learned 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, Di Cataldo et al., 2010) polyp detection, ref pulmonary emboli validation, and pathology slide classification.

Krishnan *et al.* (2011) authors studied and discussed the approaches for textural characterization of histopathological images for oral submucous fibrosis detection. The classification accuracy based on textural features for the development of computer-assisted screening of oral sub-mucous fibrosis (OSF) was improved. 71 textural features were extracted from the 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 *et al.* (2011) authors developed a knowledge-based segmentation algorithm using anisotropic diffusion and fuzzy divergence based thresholding followed by color-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. The study also observed that there is an increasing trend of BM thickness for OSFWD and OSFD compared to the 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 provided a good sensitivity of 80.16%, specificity of 100% and positive predicative accuracy of 100%.

Zhang *et al.* (2011) worked on breast cancer images with combined multiple features using the curvelet transform, statistics of completed local binary patterns (CLBP), and GLCM with a classifier Random Subspace Ensemble (RSE), with classification rate 95.22%.

Mookiah *et al.* (2011) documented an automated diagnostic techniques depending up on textural characteristics of the oral mucosal epithelium to distinguish between normal and oral submucous fibrosis (OSF). They applied this approach over 83 normal and 29 OSF histopathological image slides. They divided their approach into two parts such as Brownian motion curve for feature extraction and designing of appropriate classifiers. The distinguishing capability of the features has been validated by statistical approaches. For classifying, OSF vs normal cells, they applied error back propagation of neural networking. The sensitivity and specificity has been 100% and 85% as per Fisher's linear discriminant analysis. However, sensitivity and specificity has been 92.31% and 100% as per BPNN analysis.

Kulhalli *et al.* (2012) reported a computer aided diagnostic system and ANN detect and classify oral cancers present in Biopsy Image. The system was tested with many different types of images and found to be good.

Kasmin *et al.* (2012) extracted the features of microscopic biopsy images including area, perimeter, convex area, solidity, major axis length, orientation filled area, eccentricity, the ratio of cell and cells area, circularity, and mean intensity of cytoplasm.

Bergmeir *et al.* (2012) work on a model to extract the a variety of texture features correlation, homogeneity, energy, gray level, and HSV by using local histograms and GLCM. The quasi-supervised learning algorithm operates on two datasets, the first one having normal tissues labeled indirectly and the second contain an unlabeled collection of mixed samples of both normal and cancer tissues. He proposed method was applied to the dataset of 22,080 vectors with reduced dimensionality 119 from 132. The resulting labeling performances were compared to that of a conservative powerful supervised classifier using manually labeled ground-truth data. The experimental results showed that the supervised classifier results were calculated false positive and true positive rate as 3.5% and 95% for the same case.

Ali and Madabhushi (2012), authors worked on an unsupervised method for the segmentation and classification of cells. The procedure used constructs a tree using hierarchical clustering and then arranges the cells in a linear order by using an optimal leaf ordering algorithm which maximizes the similarity of adjacent leaves without any requirement for training examples or parameter adjustment. Performance evaluation showed the effectiveness of the proposed method in images having inconsistent staining, poor contrast, and overlapping cell.

Genc-tav and Aksoy (2012) reported categorization of histopathological images belonging to four types of carcinoma such as: cervix, prostate, breast and lung. For diagnosing cervix carcinoma, a Bayesian network is coded for constructing decision support system required in usual estimation of grade in cervical intraepithelial neoplasia (CIN). The SVM and MRF are utilized for categorization and grading of prostrate carcinoma from histopathology images of prostate cancer cells. The Bloom-Richardson (BR) grading approach and extract feature along with SVM approaches have been utilized for classifying the breast and lung carcinoma respectively.

Belsare and Mushrif (2012) represented an innovative approach for usual medical assessment of kidney cancer cells as obtained from the histopathology images utilizing shape based features. These shape derived characteristics elaborates the sharing of features as obtained from prevailing H & E staining colours in kidney cancer cells. They applied 4-class estimation performance for shape based categorisation models utilizing 10 iterations of threefold nested CV and reported overall accuracy of 77%. The finding shows that combining shape based features along with traditional histopathological images develops the estimation performances. The automatic diagnosis system resembles the pathologist's criteria for image analysis and cancer detection.

Sheha *et al.* (2012) difference between Malignant Melanoma and Melanocytic Nevi based classifications are proposed on gray level co occurrence matrix (GLCM) by using multi layer perceptron (MLP). For discrimination of melanocytic skin tumors, texture analysis can be used for high accuracy.

Kowal *et al.* (2013) compared and tested different algorithms for nuclei segmentation on a dataset of 500 images, for which reported accuracies were ranging from 96% to 100%. Filipczuk *et al. (*2013) presented a BC diagnosis system based on the analysis of cytological images of fine needle biopsies in order to discriminate the images as either benign or malignant. Four different classifiers were trained with a 25-D feature vector and a performance of 98% was reported on 737 images.

Amaral *et al.* (2013) presented a computational pipeline for automatically classifying and scoring breast cancer TMA spots mapped onto an ordinal scale used by pathologists. MLP classifier is compared with support vector machines and latent topic models for spot classification and with Gaussian process ordinal regression and linear models for scoring. They used k-NN and neural network classifier for classification accuracy 86% and 92%, respectively. Rathore, (2014) represented colon biopsy image based categorizations. They examined over 174 colon biopsy samples to categorize it with linear, radial basis function (RBF) and sigmoid SVM classifiers. They reported the classification performance near to 98.85%. Dhivya *et al.* (2014) used Support Vector Machine (SVM) Classifiers in order to detect cancer using histopathological image analysis.

Zhang *et al.* (2013) propose a cascade approach with rejection option. The first level of the cascade was used to solve the easy cases and the hard ones are sent to a second level where a more complex pattern classification system was used. The proposed method was assessed on a database proposed by the Israel Institute of Technology, which composed of 361 images (40 × magnification). On the dataset, they reported results of 97% of reliability. In another work, done by the same authors an ensemble of one-class classifiers was assessed on the same database achieving a recognition rate of 92%.

Mouelhia, (2013), authors offered Haralicks, histogram of oriented gradients (HOG), and color component based statistical moments (CCSM), features selection and extraction approaches to categorize the cancerous cells from microscopic biopsy images. Feature extracted by authors in the study are Contrast, correlation, energy, homogeneity, Haralicks, RGB, Gray Level, HSV. The outcomes were tested on 174 colon biopsy images and improved performance calculated as 98.85%.

Ondera *et al.* (2013) authors worked on an automatic system for segmenting HCC biopsy images. For pre-processing, a dual morphological grayscale reconstruction method was used to remove noise and accentuate nuclear shapes and to obtain the initial contours of nuclei a marker controlled watershed transform was applied and a snake model was used to segment the shapes of nuclei smoothly and precisely. An SVM based decision graph classifier to classify HCC biopsy images was proposed. Investigational outcomes showed that 94.54% of classification accuracy could be attained by using the SVM-based decision-graph classifier while 90.07% and 92.88% of classification accuracy could be attained by using k-NN and SVM classifiers, respectively.

Eshlaghy *et al.* (2013) used the SVM, decision tree, and multilayer perceptron artificial neural network classifiers with feature selection. The predictors used were age at diagnosis, menarche, and menopause, tumour size, a number of involved and dissected axially lymph nodes, grade and HER2. The best classifier (SVM) reported a sensitivity, specificity, and accuracy of 96%, 91%, and 94%, respectively.

George *et al.* (2014) proposed a BC based diagnosis system on the nuclei segmentation of cytological images. They used different machine learning models, such as neural networks and support vector machines (SVMs) and reported accuracy rates ranging from 76% to 94% on a dataset of 92 images. George *et al.* (2012) evaluated datasets 92 fine needle aspiration cytology (FNAC) images to classify the benign and malignant of breast tumour. The predictive ability of support vector machine (SVM) and probabilistic neural networks (PNN) are stronger than the MLP using back-propagation algorithm and learning vector quantization (LVQ).

Nguyen *et al.* (2015) proposed a method, to calculate the tubule percentage (TP), i.e., the ratio of the tubule area to the total glandular area for 353 Haematoxylin and Eosin images of the three TSs, and plot the distribution of these TP values. This plot shows the clear division among these three scores, suggesting that the proposed algorithm is useful in distinguishing images of these TSs by using a random forest classifier.

Masood *et al.* (2016), proposed a learning model for automated diagnosis of skin cancer based on histopathological image analysis. Proposed model used deep belief net and advised SVM to train the system using a combination of labeled and unlabeled data samples. To deal with the problem of limited labeled data availability this research work presented a semi advised learning model for automated recognition of skin cancer using histopathological images. Deep belief architecture is constructed using unlabeled data by making efficient use of limited labeled data for fine tuning done the classification model. In parallel, an advised SVM algorithm is used to enhance classification results by counteracting the effect of misclassified data using advised weights. The classification performance was compared with some popular methods and the proposed model outperformed most of the popular techniques including k-NN, ANN, SVM, and semi-supervised algorithms like Expectation maximization algorithm and transductive SVM-based classification model.

Some methodologies of classification used by various author for classification of histopathology images are described in table 2.3.

Table 2.3: Previous work reported in the literature for the classification of histopathological images.

S. No.	Authors (year)	Methods used for classification	Parameters used for performance measure	Accuracy (%)
1	Sinha and Ramakrishnan, 2003	NN	Accuracy on Blood cells histology images	70.6
2	Huang *et al.*, 2009	Support vector machine (SVM)	F-measure, ROC	92.88
3	Gurcan *et al.*, 2009	Support vector machine (SVM)	F-measure, specificity.	97.00
4	Di Cathaldo *et al.*, 2010	Support vector machine (SVM)	Sensitivity and specificity as well as F-Score	91.77
5	Mookiah *et al.*, 2011	Error back-propagation neural network (BPNN) and Brownian motion curve (BMC)	Sensitivity and specificity,	92.31
6	Krishnan *et al.*, 2011	Support vector machine (SVM)	Accuracy, sensitivity etc	94.07
7	He *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
8	Genctav *et al.*, 2012	Radiating gradient vector flow (RGVF)	Weighted kappa coefficient	61.00
9	Krishnan *et al.*, 2012	Brownian motion curve (BMC), and SVM classifier	Accuracy, sensitivity etc.	88.38
10	Khurd *et al.*, 2012	SVM	Accuracy	93.70%

11	Krishnan *et al.*, 2012	Radial basis probabilistic neural network (RBPNN), Sugeno fuzzy, Decision tree (DT), (K-NN), Gaussian mixture model (GMM),	Accuracy, sensitivity, and specificity	95.70
12	Rathore *et al.*, 2014	Ensemble classification based on majority voting	Accuracy, ROC and sensitivity etc.	96.86
13	Xu *et al.*, 2014	Multiple instance learning (MIL), and multiple clustered instance learning (MCIL)	Accuracy, sensitivity, ROC curve, F-measures	86.21

2.6 Conclusion:

In this literature review section, we have focused over various prominent image processing techniques. We have discussed multiple aspects that are essential in quantitative establishment of CAD based system for analysis of histopathological slides in cancer diagnosis. This chapter exhaustively illustrates of state-of-art techniques for nuclei identification, segmentation and classification, applied in different categories of cell analysis and cancer grading through histopathological images. Preprocessing, segmentation, features extraction and classification of histopathology images strives towards getting better the accuracy, precision, and computational speed of CAD methods, and reducing the amount of manual error. At last, we have identified the limitations and open challenges in existing frameworks and give overview of proposed framework with novelties. In next chapter, we will propose methods based on enhancement and segmentation for breast cancer detection in histopathology images.