CHAPTER 3: DETECTION AND CLASSIFICATION OF CANCER FROM MICROSCOPIC BIOPSY IMAGES USING CLINICALLY SIGNIFICANT FEATURES

In this chapter, a framework for automated detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features is proposed and examined. The various stages involved in the proposed methodology include enhancement of microscopic images, segmentation of background cells, features extraction, and finally the classification. An appropriate and efficient method is employed in each of the design step of the proposed framework after making a comparative analysis of commonly used method in each category. For the enhancement of the microscopic biopsy images, the contrast limited adaptive histogram equalization approach is used to highlight the details of the tissue and structures. For the segmentation of background cells, k-means segmentation algorithm is used because it performs better in comparison to other commonly used segmentation methods. In feature extraction phase, it is proposed to extract various biologically interpretable and clinically significant shape as well as morphology based features from the segmented images. These include gray level texture features, color based features, color gray level texture features, Law's Texture Energy (LTE) based features, Tamura's features, and wavelet features. Finally, the K-Nearest Neighborhood (KNN) based method is used for classification of images into normal and cancerous categories because it is performing better in comparison to other commonly used methods for this application such as fuzzy KNN and Support Vector Machine (SVM) based classifiers. The performance of the proposed framework is evaluated using well known parameters for four fundamental tissues (connective, epithelial, muscular and nervous) of randomly selected 1000 microscopic biopsy images.

3.1 Introduction

Cancer detection has always been a major issue for the pathologists and medical practitioners for diagnosis and treatment planning. The manual identification of cancer from microscopic biopsy images is subjective in nature and may vary from expert to expert depending on their expertise and other factors which include lack of specific and accurate quantitative measures to classify the biopsy images as normal or cancerous one. The automated identification of cancerous cells from microscopic biopsy images help in alleviating the above mentioned issues and provides better results if the biologically interpretable and clinically significant feature based approaches are used for the identification of disease.

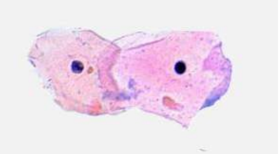
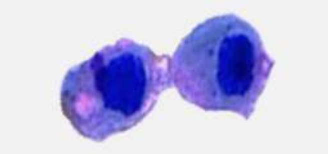
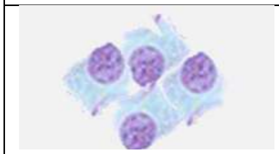
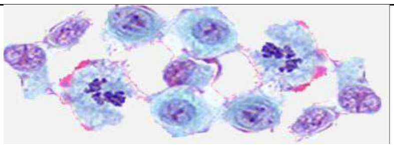
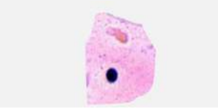
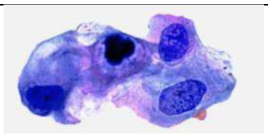
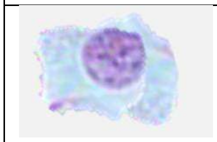
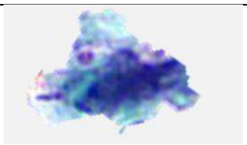
About 32% of Indian population gets cancer at some point during their life time. Cancer is one of the common disease in India which has responsibility to maximum mortality with about 0.3 million death per year (Ali, I., Wani *et al.,* 2011). The chances of getting affected by this disease are accelerated 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 from cancer are primarily in its detection and diagnosis. The selection of the treatment of cancer totally depends on its level of malignancy. Medical professionals use several techniques for detection of cancer.

These techniques may include various imaging modalities such as X-Ray, Computer Tomography (CT)-Scan, Positron Emission Tomography (PET), Ultrasound, Magnetic Resonance Imaging (MRI); and pathological tests such as urine test, blood test etc.

For accurate detection of cancer pathologists' uses histopathology biopsy images that is the 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 symptoms and indications of the disease using the microscopic biopsy images. To visualize various parts of the tissue under a microscope, the sections are dyed with one or more staining components. The main goal of staining is to reveal the components at cellular level and counter-stains are used to provide color, visibility and contrast. Hematoxylin-Eosin (H&E) are staining components that has been used by pathologists for over few decades. Hematoxylin stains cell nuclei as a blue in color while Eosin stains cytoplasm and connective tissues are of pink color. The histology (Madabhushi, A. *et al.,* 2009) is related to the study of cells in terms of structure, function and interpretations of the tissue and cells. Microscopic biopsies are most commonly used for both disease screening because of the less invasive natures. The characteristic of microscopic biopsy images has presence of isolated cells and cell clusters. The microscopic biopsy images are easier to analyze specimens compared to histopathology due to absence of non- complicated structures (Yang, L. *et al.,* 2008). The accurate manual identification of cancer from microscopic biopsy images has always been a major issue by the pathologists and medical practitioners observing at cell or tissue structure under the microscope.

In histopathology, the cancer detection process normally consists of categorizing the image biopsy into cancerous one or non-cancerous one (Gonzalez, R. C. 2009). In microscopic biopsy image analysis doctors and pathologists observe many of the abnormalities and categorizes the sample based on various characteristics of the cell nuclei such as color, shape, size, proportion to cytoplasm etc. High resolution microscopic biopsy provides reliable information for differentiating abnormal and normal tissues. The difference between normal and cancerous cells is shown in Table 3.1.

Table 3.1: Difference between normal and cancerous cells
(Liao, S., Law, M. W., & Chung, A. C. 2009)

Normal Cells	**Cancerous Cells**	**Description of Cancerous Cells**
		Large and variably shaped nuclei
		Many dividing cells and disorganized arrangements
		Variation in size and shape of nuclei
		Loss of normal feature (Shape and morphology)

For the detection and diagnosis of cancer from microscopic biopsy images, the histopathologists normally look the specific features in the cells and tissue structures. The various common features used for the detection and diagnosis of cancer from the microscopic biopsy images include shape and size of cells, shape and size of cell nuclei, and distribution of the cells. The brief descriptions of these features are given as follows (www.cancer.org):

Shape and size of the cells

It has been observed that the overall shape and size of cells in the tissues are mostly normal. The cellular structures of the cancerous cells might be either larger or shorter than normal cells. The normal cells have even shapes and functionality. Cancer cells usually do not function in a useful way and their shapes are often not even.

Size and shape of the cell's nucleus

The shape and size of the nucleus of a cancer cell is often not normal. The nucleus is decentralized in the cancer cells. The image of the cell looks like an omelet, in which the central yolk is the nucleus and the surrounding white is the cytoplasm. The nuclei of cancer cells are larger than the normal cells and deviated from the Centre of the mass. The nucleus of cancer cell is darker. The segmentation step mainly focuses on separation of regions of interests (cells) from background tissues as well as separation of nuclei from cytoplasm.

Distribution of the cells in tissue

The function of each tissue depends on the distribution and arrangements of the normal cells. The numbers of healthy cells per unit area are less in the cancerous tissues. These adjectives of microscopic biopsy images has been included in shape and morphology based features, texture features, color based features, Color Gray level Co-occurrence Matrix (GLCM), Laws Texture Energy (LTE), Tamura's features, and wavelet features are more biologically interpretable and clinically significant.

The main aim of this chapter is to design and develop a framework and a software tool for automated detection and classification of cancer from microscopic biopsy images using above mentioned clinically significant and biologically interpretable features. This chapter focuses on selecting an appropriate method for each design stage of the framework after making a comparative analysis of the various commonly used methods in each category. The various stages involved in the proposed methodology include enhancement of microscopic images, segmentation of background cells, features extraction, and finally the classification.

The rest of the chapter has been structured as follows: Section 3.2 describes the related works, Section 3.3 presents the methods and models, Section 3.4 describes the parameters setting, Section 3.5 presents the discussions of the results, and finally the Section 3.6 draws the conclusion of the work presented in this chapter.

3.2 Related Works

In recent years, few works have been reported in literature for the design and development of tools for automated cancer detection from microscopic biopsy images. Gurcan M.N. *et. al.,* (2009) presented detailed reviews on the computer aided diagnosis (CAD) for cancer detection from microscopic biopsy images. Demir, C., & Yener, B. (2005) also presented a method for automatic diagnosis of biopsy image. They presented a cellular level diagnosis system using image processing techniques. Bhattacharjee *et. al.,* (2014) presented a review on computer aided diagnosis system to detect cancer from microscopic biopsy images using image processing techniques.

Bergmeir, C. *et al.,* (2012) proposed a model to extract the texture features by using local histograms and GLCM. The quasi-supervised learning algorithm operates on two datasets, The first one having normal tissues labeled only indirectly, the second one containing an unlabeled collection of mixed samples of both normal and cancer tissues. This method was applied on the dataset of 22,080 vectors with reduced dimensionality 119 from 132. The regions having the cancerous tissues were accurately identified having true positive rate 88% and false positive rate 19% respectively by using manual ground truth data set.

Mouelhia *et al.,* (2009) used Haralick's textures features , 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 various features used in this paper are contrast, correlation, energy, homogeneity, GLCM texture features (Haralick, R. M *et al.,* ,1973) RGB, Gray Level, and HSV.

Huang, P. W., & Lai, Y. H. (2010), presented a methodology for segmentation and classification techniques for histology images based in texture features and by using SVM the maximum classification accuracy obtained is 92.8 %.

Landini, G *et al,* (2010) presented a method for morphologic characterization of cell neighborhoods in neoplastic and preneoplastic tissue of microscopic biopsy images. In this chapter, authors presented watershed transforms to compute the cell and nuclei area and other parameters. The distance measure of the neighborhood value has been used for calculating the neighborhood complexity with reference to the v-cells. The best classification has been obtained by KNN classifier is 83% for dysplastic and neoplastic classes and 58% of correct classification.

Sinha, N., Ramkrishan, *et al.,* (2003) extracted some features of microscopic biopsy images which includes eccentricity, area ratio, compactness, average values of color components, energy entropy, correlation and area of cells and nucleus. The classification accuracy obtained by Bayesian, K nearest neighbor, neural networks and support vector machine were 82.3%, 70.60%, 94.1% and 94.1% respectively.

Piuri and Scotti (2008) extracted the features of microscopic biopsy images includes area, perimeter, convex area, solidity, major axis length, orientation filled area, eccentricity, ratio of cell and nucleus area, circularity and mean intensity of cytoplasm. The KNN and Neural network classifier are used for classification accuracy 86% and 92% respectively.

In proposed chapter, a framework for automated detection and classification of cancer from microscopic biopsy images using clinically significant and biologically interpretable features is proposed and examined. For segmentation of images colour k-means based method is used. The various hybrid features which are extracted from the segmented images include shape and morphological features, GLCM texture features, Tamura features, Law's texture energy based features, histogram of oriented gradients, wavelet features, and color features. For classification purposes, k-nearest neighbor based method is proposed to be used. The efficacy of other classifiers such as SVM, Random Forest, and fuzzy k-means are also examined. For testing purposes, 2828 microscopic biopsy images available from histology database (Caicedo, J. C., Cruz, A.; Gonzalez, F. A. 2009) are used. From the obtained results, it was observed that the proposed method is performing better in comparison to other methods discussed as above. The overall summary and comparison of the proposed method and other methods are presented in Table 6 in section four of results and analysis.

3.3 Methods and Models

The detection and classification of cancer from microscopic biopsy images is a challenging task because an image usually contains many clusters and overlapping objects. The various stages involved in the proposed methodology include enhancement of microscopic images, segmentation of background cells, features extraction, and finally the classification. For the enhancement of the microscopic biopsy images, the contrast limited adaptive histogram equalization (Pisano, E. D., Zong *et al.,* 1998) approach is used and for the segmentation of background cells k-means segmentation algorithm is used. In feature extraction phase, various biologically interpretable and clinically significant shape and morphology based features are extracted from the segmented images which include gray level texture features, color based features, color gray level texture features, Law's Texture Energy (LTE) based features, Tamura's features, and wavelet features. Finally, the K-Nearest Neighborhood (KNN), fuzzy KNN and Support Vector Machine (SVM) based classifiers are examined for classifying the normal and cancerous biopsy images. These approaches are tested on four fundamental tissues (connective, epithelial, muscular and nervous) of randomly selected 1000 microscopic biopsy images. Finally, the performances of the classifiers are evaluated using well known parameters and from results and analysis, it is observed that the fuzzy KNN based classifier is performing better for the selected features set. The flowchart for the proposed work is given in Figure 3.1.

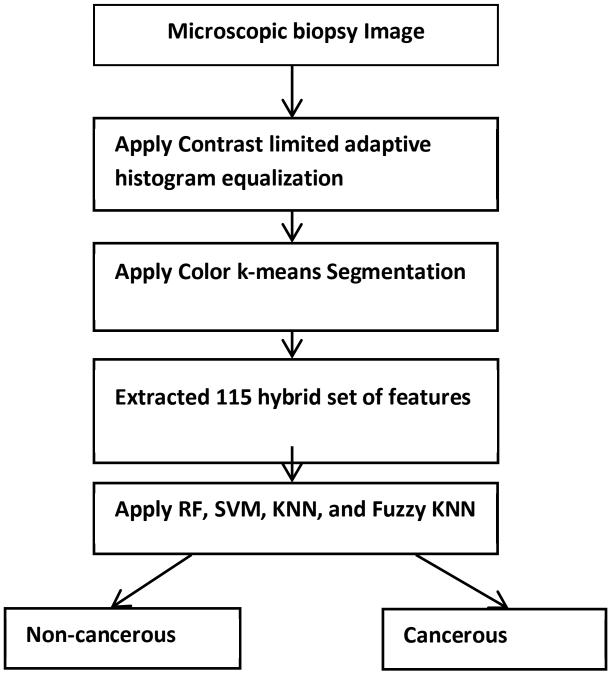

Microscopic biopsy Image
Apply Contrast limited adaptive histogram equalization
Apply Color k-means Segmentation
Extracted 115 hybrid set of features
Apply RF, SVM, KNN, and Fuzzy KNN
Non-cancerous
Cancerous

Figure 3.1: Model of automated cancer detection from microscopic biopsy images

3.3.1 Enhancements

The main purpose of the pre-processing is to remove a specific degradation such as noise reduction and contrast enhancement of region of interests. 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 (Pham, D. L. *et al.,* 2000) The contrast limited adaptive histogram equalization (CLAHE) approach is used for enhancement of microscopic biopsy images. Figure 3.2 shows the original and enhanced image using contrast limited adaptive histogram equalization.

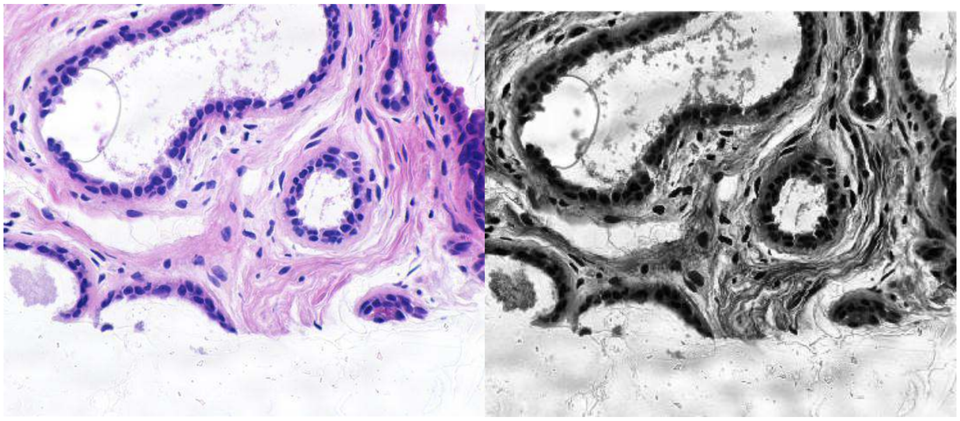


Figure 3.2: The original (left) and enhanced microscopic biopsy image with CLAHE (right) (http://www.bioimage.ucsb.edu/images/stories/BioImage/research/Benchmark/BREAST_CANCER/BreastCancerCell_dataset.tar.gz)

3.3.2 Segmentation

Several segmentation methods have been adapted for cytoplasm, cell and nuclei segmentation from microscopic biopsy images like threshold based, region based, and clustering based algorithms. However the selections of segmentation methods depend on the type of the features to be preserved and extracted. For the segmentation of ROI (region of interest), the ground truth (GT) of the images are manually cropped and created from histology dataset (http://www.bioimage.ucsb.edu/images/stories/BioImage/research/Benchmark/BREAST_CANCER/BreastCancerCell_dataset.tar.gz). The k-means clustering based segmentation algorithms is used because of the preservation of the desired information. From the obtained results through experimentation it is observed that the clustering based algorithms specifically k-means based method is the best suited for microscopic biopsy images. Figure 3.3 shows the original and k means segmented microscopic biopsy image. For testing and experimentation purpose, twenty (20) microscopic biopsy images available from histology dataset (Caicedo, J. C. *et al*., 2009) were used. These images were randomly selected for segmentation. The ground truth (GT) images are manually created by cropping the region of interest (ROI). The visual results of texture based segmentation, FCM segmentation; K-means segmentation and color based segmentation are presented in Figure 3(a) to Figure 3(d**).** Thus from the visual results obtained and reported in Figures 3(a) to Figure 3(d), it is observed that the k-means clustering based segmentation method performs better in most of the cases as compared to others segmentation approaches under consideration for microscopic biopsy image segmentation.

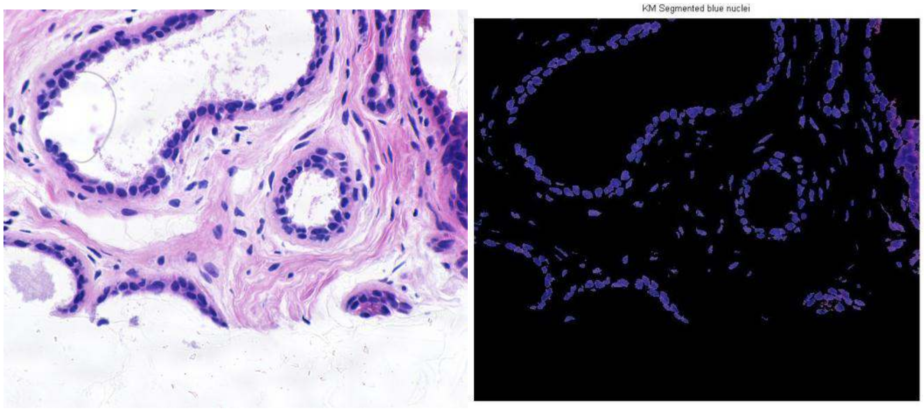


Figure 3.3: Original (left) and segmented microscopic biopsy image with K-means segmentation approach (right)

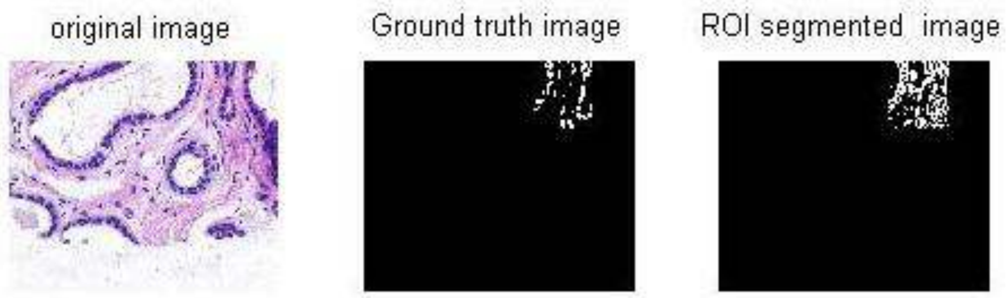


Figure 3.3(a): Original, Ground truth and ROI Segmented by Texture based segmentation

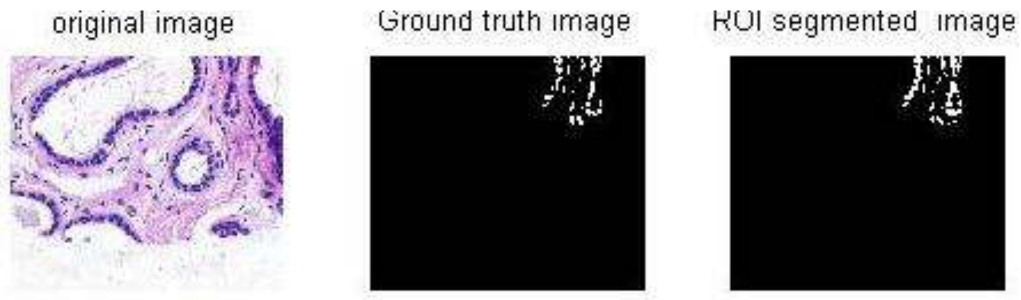


Figure 3.3(b): Original, Ground truth and ROI Segmented by FCM segmentation

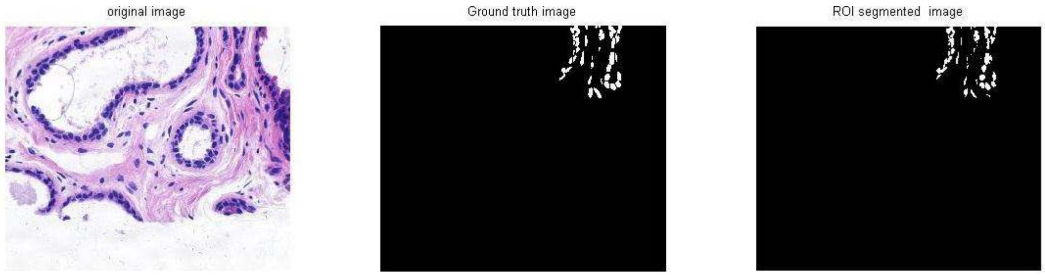


Figure 3.3(c): Original, Ground truth and ROI Segmented by k-means segmentation

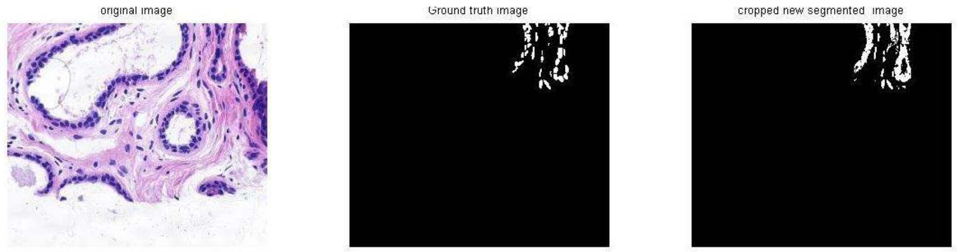


Figure 3.3(d): Original, Ground truth and ROI Segmented by color based segmentation

Finally the ROI segmented image of microscopic biopsy is compared to ground truth images for the quantitative evaluation of various segmentation approaches for all 20 sample images from histology dataset. The performance of the various segmentation approaches such as K-means (Ng, H. P. *et al.,* 2006), fuzzy c-means (Chuang, K. S. *et al*. 2006) texture based segmentation (Pal, N. R. *et al*., 1993) and color based segmentation (Wu, M. N. *et al*., 2007) were evaluated in terms of various popular parameters of segmentation measures. These parameters include accuracy, sensitivity, specificity, false positive rate (FPR), probability random index (RI), global consistency error (GCE) and variance of information (VOI).

Table 3.2 and Figure 3.4 shows the comparison of various segmentation algorithms on the basis of average accuracy, sensitivity, specificity, FPR, PRI, GCE and VOI for 20 sample images taken from histology dataset (http://www.bioimage.ucsb.edu/images/stories/BioImage/research/Benchmark/BREAST_CANCER/BreastCancerCell_dataset.tar.gz) . From the Table 3.2 and Figure 3.4, it observed that k-means, color k-means, fuzzy c-means, and texture based methods are performing better at par in terms of accuracy, specificity and PRI segmentation measures but except k-means based segmentation methods other methods are not performing better in terms of other important parameters. Only the K-means based segmentation algorithm is associated with larger value of accuracy, sensitivity, specificity, random index (RI) and smaller value of FPR, GCE and VOI in comparison to other methods and hence it is better in comparison to others. Hence, k-means based segmentation based is the only method which performing better in terms of all parameters that's why it is chosen as the segmentation method in the proposed framework for cancer detection from microscopic biopsy images.

Table 3.2: Quantitative evaluation of segmentation methods on the basis of average values of various performance metrics for a set of 20 microscopic images

Methods	**Accuracy**	**Sensitivity**	**Specificity**	**FPR**	**PRI**	**GCE**	**VOI**
color k-means	0.98	0.70	0.98	0.01	0.97	0.00	0.11
k-means	**0.98**	**0.74**	**0.99**	**0.00**	**0.98**	**0.01**	**0.10**
FCM	0.98	0.61	0.99	0.00	0.97	0.01	0.13
texture based	0.97	0.30	0.99	0.00	0.94	0.02	0.25

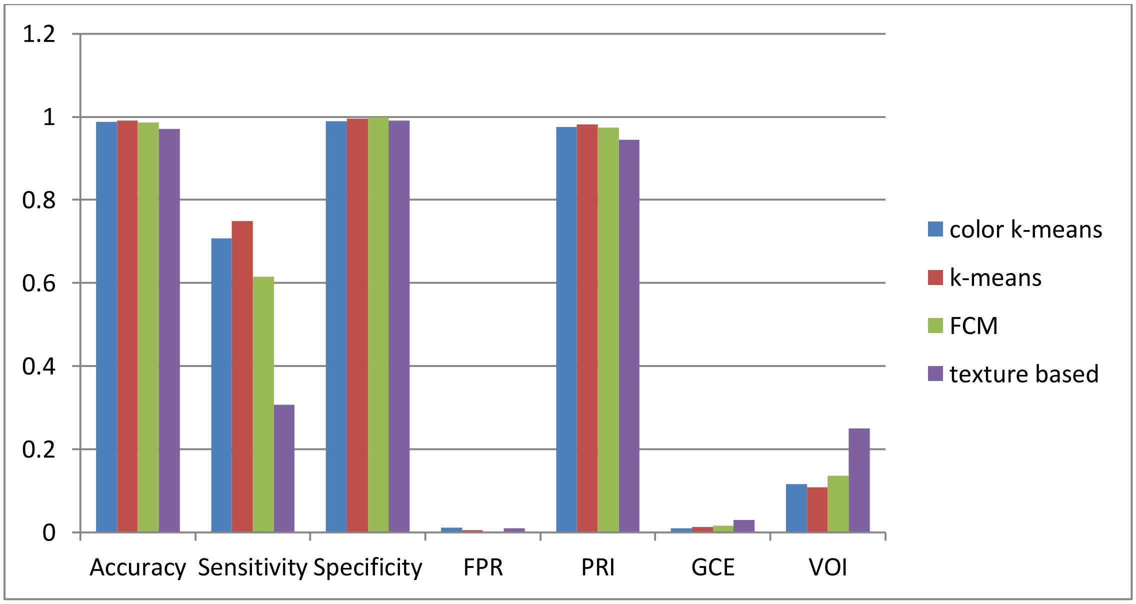


Figure 3.4: Comparisons of various segmentation methods on the basis of average Accuracy, Sensitivity, Specificity, FPR, PRI, GCE and VOI for 20 sample images from histology dataset

3.3.3 Feature Extraction

After segmentation of image features are extracted from the regions of interest to detect and grade potential cancers. Feature extraction is one of the important steps in the analysis of biopsy images. The features are extracted at tissue level and cell level of microscopic biopsy images for better predictions. 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 feature focuses on quantifying the properties of individual cells without considering spatial dependency between them. In biopsy images for a single cell, the shape and morphological, textural, histogram of oriented gradients, and wavelet features are 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.

Based on these characteristics, some important shape and morphological based features are explained as follows:

- **Nucleus area (A):** The nucleus area can be represented by nucleus region containing total number of pixels, it is shown in equation (3.1)

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} B(i, j) \tag{3.1}$$

where A=nucleus area, B is segmented image of i rows and j columns.

- **Brightness of Nucleus:** The average value of intensity of the pixels belonging to the nucleus region is known as nucleus brightness.

- **Nucleus longest diameter (NLD):** The largest circle's diameter circumscribing the nucleus region is known as nucleus longest diameter , it is shown in equation (3.2)

$$NLD = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.2}$$

  Where, x1, y1 and x2, y2 end points on major axis.

- **Nucleus shortest diameter (NSD):** This is represented through smallest circle's diameter circumscribing by the nucleus region. It is represented in equation (3.3)

$$NSD = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.3}$$

  where, x1, y1 and x2, y2 end points on minor axis.

- **Nucleus elongation:** This is represented by the ratio of the shortest diameter to the longest diameter of the nucleus region, shown in equation (3.4)

$$Nucleus\ elongation = \frac{NLD}{Perimeter} \tag{3.4}$$

- **Nucleus Perimeter (P):** The length of the perimeter of the nucleus region represented using equation (3.5).

$$P = Even\ count + \sqrt{2}(odd\ counts)\ unit \tag{3.5}$$

- **Nucleus roundness(γ):** The ratio of the nucleus area to the area of the circle corresponding to the nucleus longest diameter is known as nucleus compactness, shown in equation (3.6)

$$\gamma = \frac{A}{P} = \frac{4\pi \times Area}{P^2} \tag{ 3.6}$$

- **Solidity:** solidity is ratio of actual cell/ Nucleus area to convex hull area shown in equation (3.7)

$$Solidity = \frac{Area}{ConvexArea} \tag{3.7}$$

- **Eccentricity**: The ratio of major axis length and minor axix length is known as eccentricity and defined in in equation (3.8)

$$Eccentricity = \frac{Length\ of\ mejorAxis}{Lengthof\ \min orAxis} \tag{3.8}$$

- **Compactness**: compactness is the ratio of area and square of the perimeter it is formulated as equation (3.9)

$$Compactness = \frac{Area}{Perimeter^2} \tag{3.9}$$

There are seven set of features used for computing the feature vector of microscopic biopsy images explained as follows:

Texture Features (F1-F22):

Autocorrelation, Contrast, Correlation, Correlation, Cluster Prominence, Cluster Shade, Difference variance, Dissimilarity, Energy, Entropy, Homogeneity, Homogeneity, Maximum probability, Sum of squares, Sum average , Sum variance , Sum entropy , Difference entropy, Information measure of correlation1, Informaiton measure of correlation2, Inverse difference (INV), Inverse difference normalized (INN) and Inverse difference moment normalize are major texture features can be calculated using equations of the texture features.

Morphology and shape Feature (F23-F32): In paper (Chaudhuri, B. B. *et al.* 1988) authors describe the shape and morphology features. The considered shape and morphological features in this chapter are Area, Perimeter, Major Axis Length, Minor Axis Length, Equivalent Diameter, Orientation, Convex Area, Filled Area, Solidity, and Eccentricity.

Histogram of oriented gradient (HOG) (F33 – F68): Histogram of Oriented Gradient is one of the good features set to deification of the objects (Kong, J. *et al.,* 2009). In our observation it will be included for better and accurate identification of objects presents in microscopic biopsy images.

Wavelet Features (F69-100): Wavelets are small wave which is used to transforms the signals for effective processing. The wavelets are useful in multi resolution analysis of microscopic biopsy images because they are fast and give better compression as compared to other transforms. The Fourier transform converts a signal into a continuous series of sine waves, but the wavelet precedes it in time and frequency both. This account for the efficiency of wavelet transforms. Daubechies wavelets have been used because it has fractal structures and it is useful in the case of microscopic biopsy images. In this chapter mean, entropy, energy, contrast homogeneity and sum of wavelet coefficients are taken in the consideration.

Color features (F101-F106): The components of these models are hue, saturation and value (HSV) . This is represented by the six sided pyramids, the vertical axis behaves as brightness, the horizontal distance from the axis represents the saturation and the angle represents the hue. Here, mean and standard deviation of HSV components are taken as features.

Tamura's Features (F107-F109): Tamura's features are computed on the basis of three fundamental texture features contrast coarseness and directionality. (Madabhushi, A. *et al.,* 2009). Contrast is the measure of variety of the texture pattern. Therefore, the larger blocks that makes up the image, has a larger the contrast. It is affected by the use of varying black and white intensities (Kong, J. *et al.,* 2009). Coarseness is the measure of granularity of an image , thus coarseness can be represented using average size of regions that have the same intensity . Directionality is the measure of directions of the grey values within the image (Jain, A. K. 1989).

Laws Texture Energy (LTE) (F110- F115): These features are texture description features which mainly used average gray level, edges, spots, ripples and wave to generate vectors of the masks. Laws mask represented by the features of an image without using frequency domain. Laws significantly determined that several masks of appropriate sizes were very instructive for discriminating between different kinds of texture features presents in the microscopic biopsy images. Thus its classified samples based on expected values of variance-like square measures of these convolutions, called texture energy measures. The LTE mask method is based on texture energy transforms applied to the image classification used to estimate the energy within the pass region of filters .

Table 3.3 provides the distribution of name of the feature type the number of features selected for the classification of microscopic biopsy images.

Table 3.3: The distribution of various features extracted from images and their ranges

Name of features	**Number of features (range F1-F115)**
Texture Features	22 (F1-F22)
Morphology and shape Feature	10 (F23-F32):
Histogram of oriented gradient (HOG)	36 (F33 – F68)
Wavelet Features	32 (F69-100)
Color features	6 (F101- F106)
Tamura's Features	3 (F107- F109)
Laws Texture Energy	16 (F110- F115)

3.3.4 Classification

The classification of microscopic biopsy images is most challenging task for automatic detection of cancer from microscopic biopsy images. Classification might provide 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 neighbor classifiers, support vector machine (SVM) and random forest (RF). Supervised machine learning approaches are used for the classification of microscopic biopsy images. There are various steps involves in the supervised learning approaches. First step is to prepare the data(feature set) , the second step is to choose an appropriate algorithms, the third step is to fit a model , the fourth step is to train the fitted model and then the final step is to use fitted model for prediction. The K-Nearest Neighborhood (KNN), fuzzy KNN and Support Vector Machine (SVM) , and Random Forest classifiers are used for classifying the normal and cancerous biopsy images.

3.4 Parameters Setting

The proposed methodologies were implemented with MATLAB 2013b, on data set of digitized at 5× magnification on PC with 3.4 GHz Intel Core i7 processor, 2GB RAM and windows7 platform.

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. 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. The brief description of dataset used for identification of cancer from microscopic biopsy images are provided in Table 2.4.

The proposed methodology for detection and diagnosis of cancer detection from microscopic biopsy images consists of the stages of images enhancement, segmentation, feature extraction, and classification.

For enhancement of microscopic biopsy images, the contrast-limited adaptive histogram equalization (CLAHE) (Pisano, E. D. *et al.,* 1994) was used as they are better able to highlight the regions of interests in the images as tested through experimentation.

To better preserve the desired information in microscopic biopsy images during segmentation process, the various clustering and texture based segmentation approaches were examined. For microscopic biopsy images it is required to discover as much as possible the nuclei information in order to make reliable and accurate detection and diagnosis based on cells and nuclei parameters. From results and analysis presented in section-3.3, k-means segmentation algorithm was used for segmenting the microscopic biopsy images as it performs better in comparison to other methods. During segmentation process of k-means clustering method, the number of clusters k were set to k=3. Further, to find the center of the clusters, squared Euclidean distance measures are used as similarity measures.

In feature extraction phase, various biologically interpretable and clinically significant shape and morphology based features were extracted from the segmented images which include gray level texture features (F1-F22), shape and morphology based features (F23-F32), histogram of oriented gradients (F33-F68) , wavelet features (F69-F100) ,color based features (F101-F106), Tamura’s features (F107-F119) and Law’s Texture Energy (F110-F115) based features. Finally a 2-D matrix of 2828×115 feature matrix was formed using all the feature set, where 2828 are the number of microscopic images in the dataset and 115 are the total number of features extracted.

Randomly selected 1000 data/samples were used for testing various classification algorithms. The 10-fold cross validation approach was used to partition the data in training and testing sets. Thus 900 data/samples were used for training purposes and 100 data/samples were used for testing purposes. The K-nearest neighbor (KNN) is simple classifier in which a feature vector is assigned. For KNN classification the numbers of nearest neighbor (k) were set to 5, and Euclidean distance matrix and the 'nearest' rule to decide how to classify the sample were used. The proposed method was also tested by using support vector machine (SVM) based classifier for linear kernel function with 10-fold cross validation methods. In SVM classification model, the kernel's parameters, and soft margin parameter C plays vital role in classification process, the best combination of C and γ were selected by a grid search with exponentially growing sequences of C and γ. Each combination of parameter choices was checked using cross validations (10 fold), and the parameters with best cross validation accuracy were selected. For SVM's linear kernel function, Quadratic programming (QP) optimization parameter was used to find separating hyper plane. In the case of random forest the value by defaults 500 trees and mtry=10.

The performance of classifiers were calculated using confusion matrix of size 2×2 matrix and the value of TP, TN, FP, and FN were calculated. The performance parameters accuracy, sensitivity and specificity were calculated using Equations (3.14) - (3.21).

3.5 Discussions of Results:

Table 3.4 shows classification results of the proposed framework for four different tissues of microscopic biopsy images containing cancer and non-cancer cases tested using four popular classifiers like k-nearest neighbor, SVM, fuzzy KNN and Random Forest.

From Table 3.4 and Figure 3.5 following observations are made for sample test cases containing Connective Tissues:

- For the identification of cancer from biopsy images of Connective Tissues in the case of KNN, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.921909, 0.940164, 0.819922, 0.880263, 0.759395 and 0.717455 respectively.
- For the identification of cancer from biopsy of Connective Tissues in the case of SVM, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.89245, 0.888438, 0.948297, 0.918756, 0.538314 and 0.55879 respectively.
- For the identification of cancer from biopsy of Connective Tissues in the case of Fuzzy KNN, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.787879, 0.867476, 0.370074, 0.618789, 0.356613 and 0.231013 respectively.
- For the identification of cancer from biopsy of Connective Tissues, in the case of Random forest classifier, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.907245, 0.993668, 0.493996, 0.743832, 0.647373 and 0.642137 respectively.

From Table 3.5 and Figure 3.6 following observations are made for sample test cases containing Epithelial Tissues:

- For the identification of cancer from biopsy images of Epithelial Tissues in the case of KNN, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.884727, 0.916446, 0.801733, 0.859435, 0.795319 and 0.71626 respectively.

- For the identification of cancer from biopsy of Epithelial Tissues in the case of SVM, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.796998, 0.7851, 0.898525, 0.842279, 0.472804 and 0.4587 respectively.
- For the identification of cancer from biopsy of Epithelial Tissues in the case of Fuzzy KNN, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.665834, 0.76465, 0.407057, 0.585984, 0.401181 and 0.17053 respectively.
- For the identification of cancer from biopsy of Epithelial Tissues, in the case of Random forest classifier, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.849306, 0.966243, 0.555332, 0.760788, 0.675868 and 0.609494 respectively.

From Table 3.6 and Figure 3.7 following observations are made for sample test cases containing Muscular Tissues:

- For the identification of cancer from biopsy images of Muscular Tissues in the case of KNN, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.897321, 0.923277, 0.650761, 0.787092, 0.543009 and 0.49783 respectively.
- For the identification of cancer from biopsy of Muscular Tissues in the case of SVM, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.884379, 0.886718, 0.786303, 0.83681, 0.263764 and 0.320547 respectively.
- For the identification of cancer from biopsy of Muscular Tissues in the case of Fuzzy KNN, the average value of accuracy, specificity,

sensitivity, BCR, F-measure and MCC are 0.614958, 0.672503, 0.535894, 0.604364, 0.538571 and 0.208941 respectively.

- For the identification of cancer from biopsy of Muscular Tissues, in the case of Random forest classifier, the accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.889878, 0.995023, 0.193145, 0.594084, 0.313309 and 0.37318 respectively.

From Table 3.7 and Figure 3.8 following observations are made for sample test cases containing Nervous Tissues:

- For the identification of cancer from biopsy images of Nervous Tissues in the case of KNN, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.861763, 0.880866, 0.835733, 0.858482, 0.834116 and 0.716492 respectively.
- For the identification of cancer from biopsy of Nervous Tissues in the case of SVM, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.769545, 0.723056, 0.946068, 0.834923, 0.630126 and 0.552038 respectively.
- For the identification of cancer from biopsy of Nervous Tissues in the case of Fuzzy KNN, the accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.808453, 0.882722, 0.242776, 0.562835, 0.225886 and 0.11837 respectively.
- For the identification of cancer from biopsy of Nervous Tissues, in the case of Random forest classifier, the average value of accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.843102, 0.92827, 0.723262, 0.825766, 0.792403 and 0.676888 respectively.

From above discussions for all four categories of test cases, it is observed that the KNN is performing better in comparison to other classifiers for the identification of cancer from biopsy images of Nervous Tissues.

From all above observations, it is concluded that the KNN classifier is producing better results in comparison to other methods for the case of biopsy images of connective tissues. The maximum values of the accuracy, sensitivity, and specificity are 0.9552, 0.9615 and 0.9543 respectively. The k-nearest neighbor classifier is also performing better for all cases as well as discussed above. Table 3.9 gives a comparative analysis of the proposed framework with other standard methods available in literature. From this, Table 3.5, it can be observed that the proposed method is performing better in comparison to all other methods.

Table 3.4: Performance analysis of classifiers on connective Tissue

Connective Tissue						
Classification Methods	**Accuracy**	**Specificity**	**Sensitivity**	**BCR**	**F-measure**	**MCC**
RF	0.907245	0.993668	0.493996	0.743832	0.647373	0.642137
SVM	0.89245	0.888438	0.948297	0.918756	0.538314	0.55879
FYZZY KNN	0.787879	0.867476	0.370074	0.618789	0.356613	0.231013
KNN	**0.921909**	**0.940164**	**0.819922**	**0.880263**	**0.759395**	**0.717455**

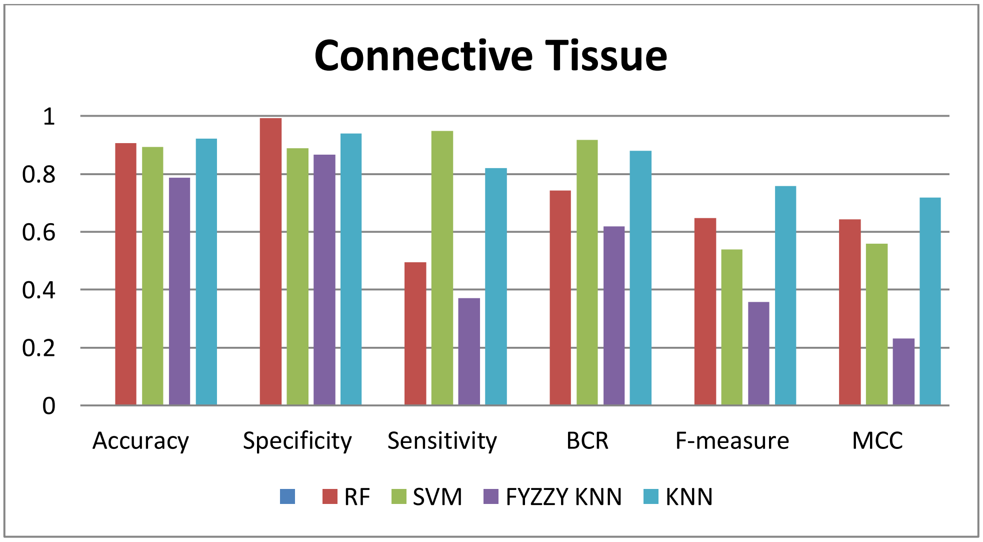


Figure 3.5: Performance analysis of classifiers on connective Tissue

Table 3.5: Performance analysis of classifiers on Epithelial Tissues

Epithelial Tissues						
Classification Methods	**Accuracy**	**Specificity**	**Sensitivity**	**BCR**	**F-measure**	**MCC**
RF	0.849306	0.966243	0.555332	0.760788	0.675868	0.609494
SVM	0.796998	0.7851	0.898525	0.842279	0.472804	0.4587
FYZZY KNN	0.665834	0.76465	0.407057	0.585984	0.401181	0.17053
KNN	**0.884727**	**0.916446**	**0.801733**	**0.859435**	**0.795319**	**0.71626**

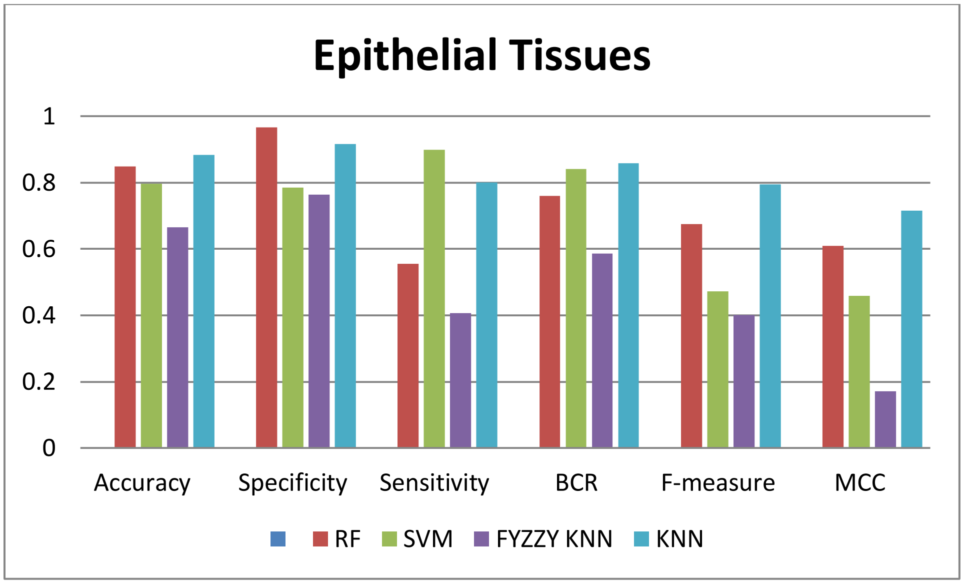


Figure 3.6: Performance analysis of classifiers on Epithelial Tissues

Table 3.6: Performance analysis of classifiers on Muscular Tissues

Muscular Tissues						
Classification Methods	**Accuracy**	**Specificity**	**Sensitivity**	**BCR**	**F-measure**	**MCC**
RF	0.889878	0.995023	0.193145	0.594084	0.313309	0.37318
SVM	0.884379	0.886718	0.786303	0.83681	0.263764	0.320547
FYZZY KNN	0.614958	0.672503	0.535894	0.604364	0.538571	0.208941
KNN	**0.897321**	**0.923277**	**0.650761**	**0.787092**	**0.543009**	**0.49783**

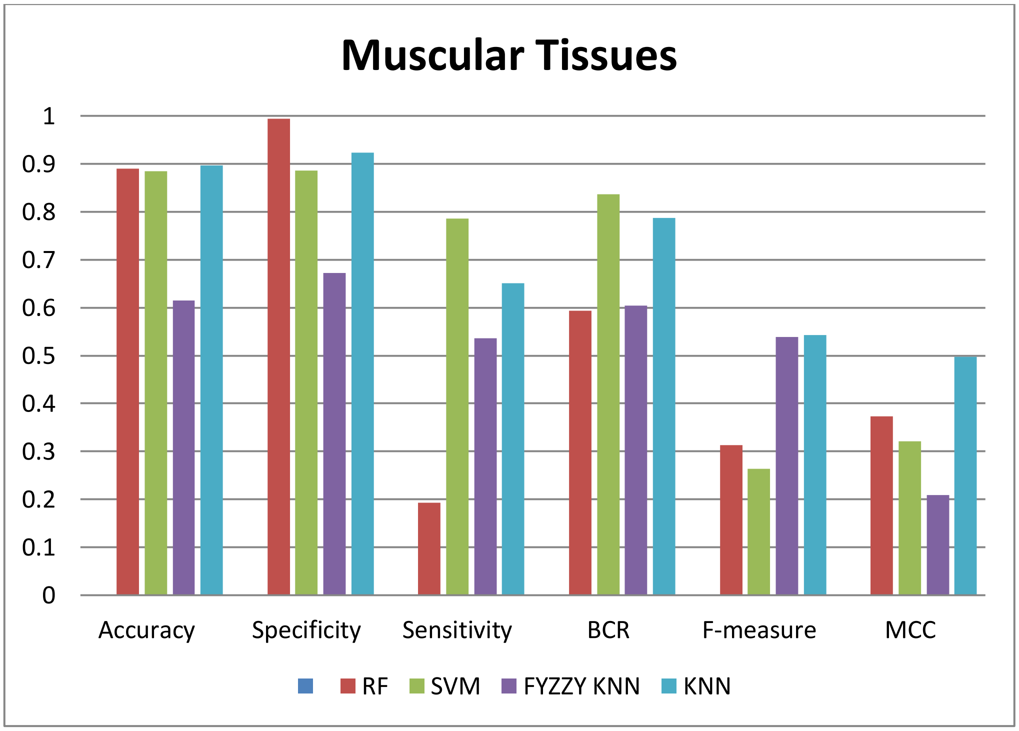


Figure 3.7: Performance analysis of classifiers on Muscular Tissues

Table 3.7: Performance analysis of classifiers on Nervous Tissues

Nervous Tissues						
Classification Methods	**Accuracy**	**Specificity**	**Sensitivity**	**BCR**	**F-measure**	**MCC**
RF	0.843102	0.92827	0.723262	0.825766	0.792403	0.676888
SVM	0.769545	0.723056	0.946068	0.834923	0.630126	0.552038
FYZZY KNN	0.808453	0.882722	0.242776	0.562835	0.225886	0.11837
KNN	**0.861763**	**0.880866**	**0.835733**	**0.858482**	**0.834116**	**0.716492**

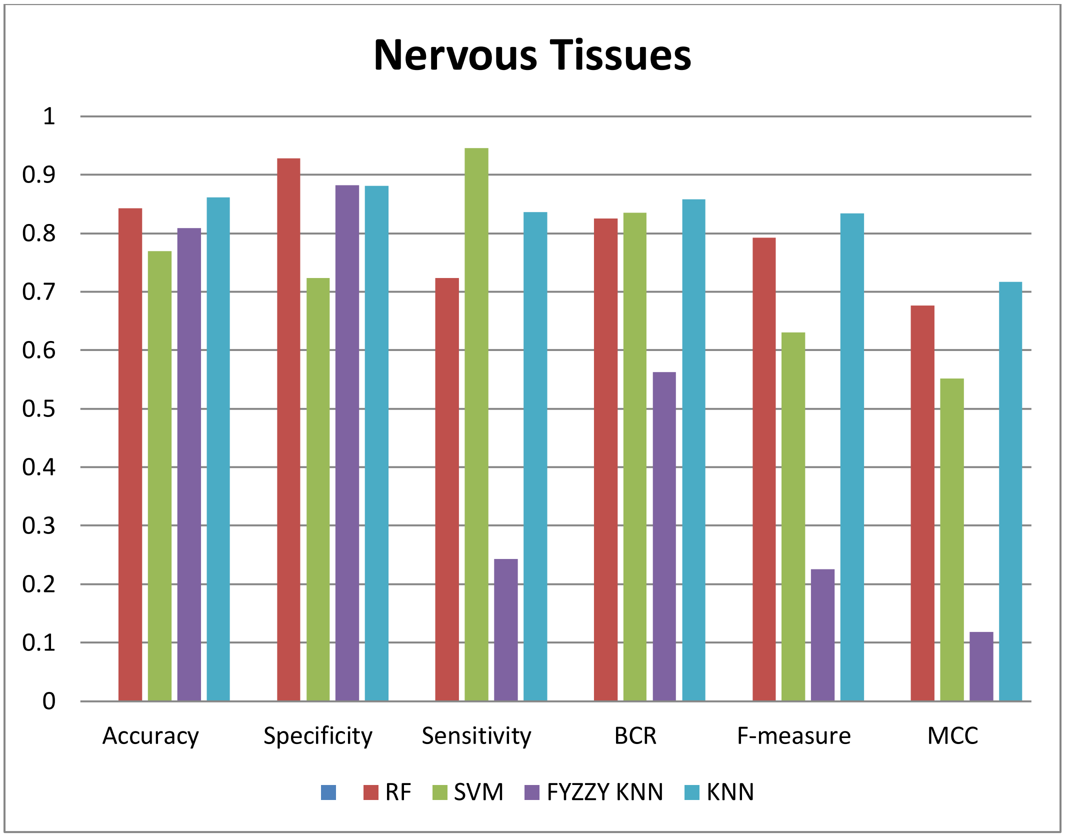


Figure 3.8: Performance analysis of classifiers on Nervous Tissues

Table 3.8: The Comparison of the proposed method with other standard methods

Authors (Year)	**Feature Set used**	**Methods of classification**	**Parameters used (%)**	**Data Set used**
P.W. Huang *et. al,.(2009)*	Texture features	Support Vector Machine (SVM)	Accuracy =92.8	1000×1000,4000×3000 and 275×275 HCC biopsy images
S. Di Cathaldo *et al.,* (2010)	Texture and morphology	Support Vector Machine (SVM)	Accuracy =91.77	Digitized histology lung cancer IHC tissue images
HE Lei *et. al,* (2012)	Shape, morphology and texture	Artificial Neural Network (ANN), (SVM),	Accuracy =90.00	Digitized histology images

MR Mookiah *et al.,* (2011)	Texture and morphology	Error Back-Propagation Neural Network (BPNN)	Accuracy =91.43,
			Sensitivity=92.31,
			Specificity=82
				83 normal and 29 OSF images

Krishnan *et al.,* (2011)	HOG, LBP, LTE	LDA	Accuracy =82	Normal -83 OSFWD-29
M. Muthu Rama Krishnan *et al*, (2011)	HOG, LBP, LTE	Support Vector Machine (SVM)	Accuracy =88.38	Histology images Normal-90 OSFWD-42 OSFD-26

Caicedo, J. C. *et at.*, (2011)	Bag of features	Support Vector Machine (SVM)	Sensitivity=92
			Specificity=88
				2828 Histology images

Sinha and Ramakrishan *et at.*, (2003)	Texture and statistical features	KNN	Accuracy=70.6	Blood cells histology images
The proposed approach	**Texture, shape and morphology, HOG, wavelet color, Tamura's and LTE**	**KNN**	**Average:** **Accuracy =92.19** **Sensitivity=94.01,** **Specificity=81.99** **BCR=88.02** **F-measure=75.94** **MCC=71.74**	**2828 Histology images**

3.6 Conclusions

An automated detection and classification procedure was presented for detection of cancer from microscopic biopsy images using clinically significant and biologically interpretable set of features. The proposed analysis was based on tissues level microscopic observations of cell and nuclei for cancer detection and classification. The various classification approaches tested were K-Nearest Neighborhood (KNN), fuzzy KNN, Support Vector Machine (SVM), and random forest based classifiers. From Table 3.8 we are in position to conclude that, KNN is best suited classification algorithm for detection of Non-cancerous and cancerous microscopic biopsy images containing of all four fundamental tissues. SVM provides average results for all the tissues types but not better than KNN. Fuzzy KNN is comparatively less good classifier. RF classifier provides very low sensitivity and down accuracy rate as compared to KNN classifier for this data set. Hence, from experimental results, it was observed that KNN classifier is performing better for all categories of test cases present in the selected test data. These categories of test data are connective tissues, epithelial tissues, muscular tissues, and nervous tissues. Among all categories of test cases, further it was observed that the proposed method is performing better for connective tissues type sample test cases. The performance measures for connective tissues dataset in terms of the average accuracy, specificity, sensitivity, BCR, F-measure and MCC are 0.921909, 0.940164, 0.819922, 0.880263, 0.759395 and 0.717455 respectively.