

Classification of Cancerous and Non Cancerous Cells in H & E Breast Cancer Images Using Structure Descriptors

B. Lakshmanan^{1*}, S. Anand², S. Vijay Gokul³

¹Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India

²Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.

³Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India.

*lakshmanan@mepcoeng.ac.in

Article History Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Computer-assisted earlier detection and diagnosis of breast cancer are possible by analyzing morphological features in histopathology images. In this article, we propose computer-aided tool for accurate detection of the abnormal cells using shapes and morphological based features extracted from the segmented region of the input image. Different segmentation techniques have been performed and their performance is compared to find out the best segmentation algorithm for further processing i.e., classification. Morphological descriptor such as major and minor-axis length, cell area, perimeter and eccentricity are extracted from the segmented region of the image. By analyzing the cell mass and contour of cells in the given input image, cells are classified as cancerous and non-cancerous. The results are validated using benchmark dataset images taken from Mitosis Atypia14 grand challenge in which 180 images are taken for training and 120 images for testing. The proposed method provides improved F-score of 94.12% compared with other previously proposed frameworks.

Keywords: Histopathology Images, Segmentation, Cancer Cells, Non-cancerous Cells.

1. Introduction

Cancer is a serious health issue worldwide and also results in a high mortality rate. Studies conducted by the International Agency for Research on Cancer of the World Health Organization estimate that there will be new cases of 27 million before 2030 and death of 8.2 million were caused by cancer in recent years [1]. In 2018 an average of 1 million cancers affected cases and also six hundred thousand deaths were registered in the US [2]. In developing countries like India, the study says the country reports a high death rate in cancer disease and stands the second position worldwide in terms of the maximum mortality rate with 0.3 million deaths for every year. The major categories of cancer in human include breast, skin, lungs, cervix and oral cancer. Breast cancer is considered as the most common cancer type among women worldwide and it ranks next to lung cancer in case of death rate [3].

Different imaging modalities for cancer diagnosis include mammography, magnetic resonance imaging (MRI), Positron emission tomography (PET) and tissue biopsy method. Breast Cancer detection and diagnosis through Digital Histopathology images get much attention today by the physician and medical image researchers for the accurate progression detection of the disease with the help of recent advancements in machine learning algorithms. The diagnosis from histopathology images remains the “gold standard” in medical diagnostic procedure for various diseases including cancer disease [6]. In tissue biopsy method, a small tissue region or sample or mass is extracted from the affected region using a medical procedure like surgery or fine needle aspiration (FNA), and followed by prepare a slide [7]. The manual annotation in H & E slide is performed by a pathologist in examines the cancerous and non-cancerous cells in the image and their role in decision making will helpful for cancer prognosis. The manual analysis of huge biopsy slides by the pathologist is a labor-intensive and challenging work. This can be overcome by advancements in digital pathology, the autonomous detection of abnormal structural patterns in a whole slide image (WSI) will provide valuable insight to the pathologist today. Hematoxylin and Eosin (H & E) stains [7] are used in slide preparation, here nuclei turn into blue colour and cytoplasm converts to pink like colour due to Hematoxylin and Eosin stains respectively. The stain variant images are shown in Figure 1(a)-(c).

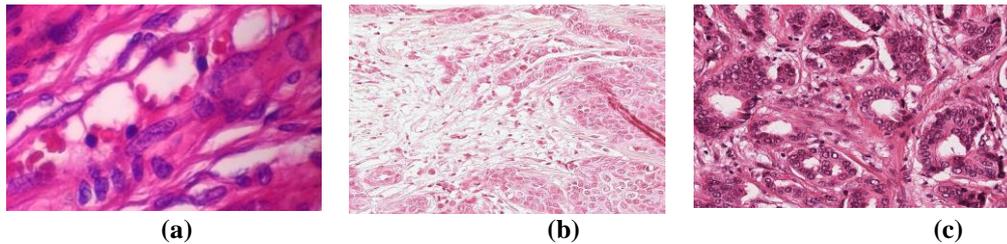


Figure 1 (a)-(c). Histopathology Images with Various Stains Collected from MITOS-ATYPIA14 Contest Dataset [28]

Tumour cells duplicate at a faster rate compared to other cells. Cells multiply rapidly and turn into a tumour during the proliferation stage that may be a malignant tumour and benign tumour [8]. The study of cellular morphological structure in the digitized histopathology image is considered as the best method in the detection of malignant cells and the progression rate of cancer. The highlight of morphological features based approach is that these features will describe the shapes of the tumour cells. Cell morphological analysis is an important prognostic factor that helps to investigate the cell behaviour and growth progression of the microscopic cells in various stages [9]. The scientific advancements in developing Computed Aided Diagnostic (CAD) tool which helps to locate and detect the abnormal cellular structure in histopathology image. Early detection and diagnosis system through CAD tools will be more helpful to reduce the mortality rate.

Aswathy et al.,[10] presented various techniques for nuclei detection, segmentation and classification. They suggest the various analytic tools that will help for accurate detection of cancer. Wang et al.,[11], combine the multi-scale region-growing algorithm and wavelet decomposition method to locate ROIs. They also used the morphology operation augmented curvature scale space corner detection algorithm for cell segmentation. Basavanahally et al.,[12] proposed the use of combine color gradient based active contour model (ACM) and hierarchical normalized cut approach for better segmentation results. In Dundar et al.,[13], Gaussian mixture model and expectation-maximization algorithm have been used for nuclei segmentation. In Tosun and Gunduz-Demir[14], graph run-length matrices used for image segmentation has been discussed. Their approach results in good accuracy rate for cell segmentation.

Veta et al., [15] applied a marker-controlled based watershed algorithm for the segmentation of stained histopathology images. Their approach achieved the accuracy rate of 81%. In Jain et al.,[16], combine ACM and neural network-based classifier for classifying cells into a malignant cell and benign cell. Their method results in classifier accuracy of 83.47%. Niwas et al.,[17] used a complex wavelet transform for the feature extraction. The features specify the chromatin structure of the tumour cell. They used k-NN classifier as a training model and achieved good classification rate. In [18], contrast restricted adaptive based histogram equalization algorithm for histopathology images has been discussed. Nie et al.,[19] propose a method to construct feature gray level co-occurrence matrix. They used a feature matrix for model training. Wikinson et al.,[20] proposed the adaptive threshold method for image segmentation in microbes images. Their approach is appropriate for a noisy image. He Le et al [21] applied an algorithm for thresholding based Gaussian mixture modelling for segmentation task in histology stained images. In Demir et al [22], digital image processing techniques in tissue and cellular level diagnosis of biopsy images has been discussed.

The main objective of the article is to investigate a few algorithms for nuclei segmentation and classification of cells into cancerous and non-cancerous. In this article, various techniques in digital image processing adopt on histopathology images and its results have been analyzed. Classification of cells into a malignant category and benign category has performed in three stages: Pre-processing of the input image, nuclei cell segmentation and feature extraction.

The article is structured as follows: Section 2 gives materials and methods. Section 3 discusses the experimental results and discussions. Finally, Section 4 records the conclusion of the work along with future possibilities explained in Section 5.

2. Materials and Methods

Data Set

Histopathology breast cancer images have been taken from ICPR MITOSIS dataset 2014[29]. The dataset consists of 1696 high-power fields (HPFs) at different magnification factor. In our experimental practice, we consider the HPF with a 40X magnification factor. Each HPF image is of 1539 x 1376 pixels size. We have 1200

HPF images with 749 labelled mitotic cells for training and 496 HPFs for testing. In training dataset, the abnormal cells and normal cells are annotated by a well experienced pathologist.

Nuclei Detection and Segmentation

In the histopathology image, nuclei morphological structure varies in different stages of cancer progression. By examining the morphological cell structure in histopathology image, we can obtain a conclusive idea about the progression rate of cancer growth. During histopathology image analysis, information about size, the shape of the tumour and other nuclei cell structure can be obtained. Therefore, it necessitates the need for detecting and segmentation of nuclei structure in an image and this process is an important step in cancer grading. Basic segmentation techniques such as Binarization, Otsu thresholding and watershed transform is used to segment the nuclei. The proposed framework consists of pre-processing stage and nuclei segmentation by standard algorithms. The block diagram of our work is shown in Figure 2.

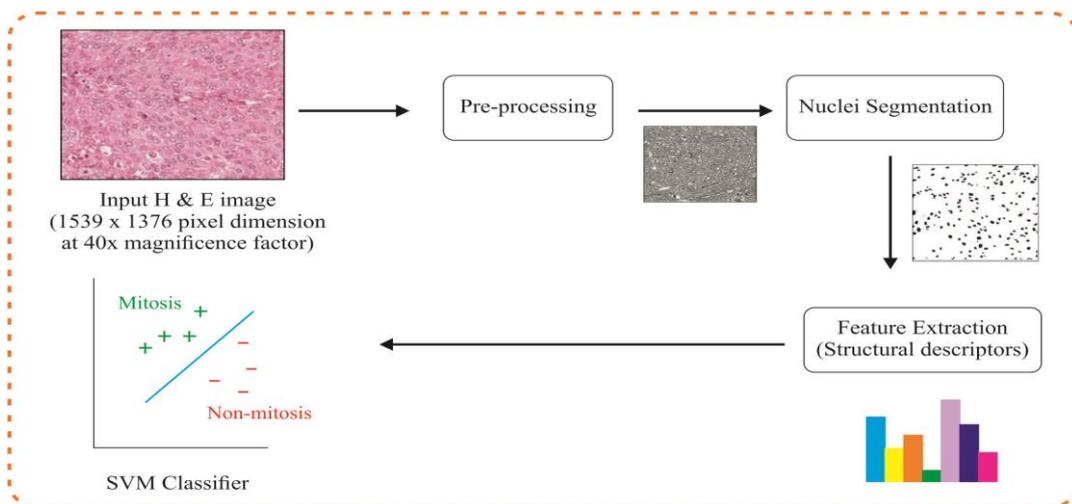


Figure 2. Block Diagram of the Proposed Work

Preprocessing

In this work, Let us consider an image is a function and it is denoted by I ,

$$I : X \rightarrow [0,1]^c \quad (1)$$

where $X = [[0; p - 1]] \times [[0; q - 1]]$ are the pixels, p and q denotes numbers of rows and columns, and c is the number of color channels. The input image as shown in Figure 3, I are preprocessed using a median filter to remove any form of graininess noise. The median filter considers each pixel in the image I and replaces it with the median of the neighbourhood pixel values. A Grayscale image is given as input and noise filtered image is returned as output as shown in Figure 4.

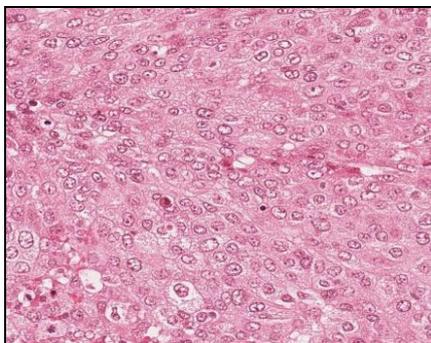


Figure 3. Input H & E Image (1539 x 1376 Pixel Dimension at 40x Magnification Factor)

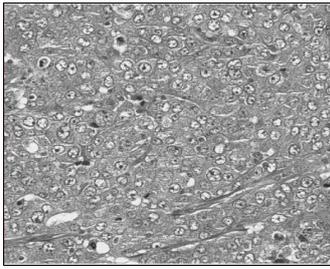


Figure 4. Preprocessed Image (Median Filter)

Segmentation using Binarization

In this technique, the given preprocessed histopathology image is converted into a binary image.

The pre-processed output is given as input and it is transformed into a bi-level image using the threshold, T. In this segmentation procedure, the all pixels in an image are assigning to either 0's or 1's with the help of T.

$$g(x, y) = \begin{cases} 1 & \text{if } I(x, y) \geq T \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

For the images that we have taken, the threshold value of 0.6 produced better binarization after various trails. So, in our work T value is fixed as 0.6. Binarization technique is performed as per equation (2). Thus the nuclei cells become black in color (pixel value of 1) and the background region other than the cells become white in color (pixel value of 0).

Otsu's Thresholding based Segmentation

Otsu's technique [28] will perform image thresholding via clustering method and it reduces grayscale image to a binary image. In this method, the assumption is that the image contains two class distributions of pixels i.e., foreground class pixels and background class pixels. The separation of two classes is performed by estimate the optimum threshold value and check whether the intra-class variance is minimal, or equivalently their inter-class variance is maximal. As the variance sum and within-class variance are known, between class variance can easily be calculated. Run through a whole range of t varies from 1 to maximum pixel value in an image and select the t value that maximizes the between-class variance. Otsu algorithm [28] is summarized in Algorithm 1.

Algorithm 1: Otsu Algorithm [28]

Input: Preprocessed image

Output: Image into two classes

1. Minimize $\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t)$
2. Compute the class probabilities

$$q_1(t) = \sum_{i=1}^t p(i)$$

$$q_2(t) = \sum_{i=t+1}^I p(i)$$

Where p represents the histogram of an image.

3. Estimate the average of each cluster

$$\mu_1(t) = \frac{\sum_{i=1}^t ip(i)}{q_1(t)} \text{ and } \mu_2(t) = \frac{\sum_{i=t+1}^I ip(i)}{q_2(t)}$$

Where μ_1 and μ_2 are mean of first cluster and second cluster respectively.

q_1 and q_2 are class probability of first class and second class respectively.

4. Compute the individual class variance

$$\sigma_1^2(t) = \sum_{i=1}^I [i - \mu_1(t)]^2 \frac{p(i)}{q_1(t)}$$

$$\sigma_2^2(t) = \sum_{i=t+1}^I [i - \mu_2(t)]^2 \frac{p(i)}{q_2(t)}$$

5. Find between class variance

Segmentation Using Watershed Algorithm

In this method, the segmentation process usually starts from specific pixels in an image called markers and progressively floods the neighbouring regions of markers, called catchment basin (CB). Watershed lines are defined as maximum altitude lines where CB is separated topographically by these lines from nearby CBs. In this algorithm, it classifies every point of a topographic plane as either belonging to the CB related to one among the local minima or the watershed line. The detailed algorithm for watershed segmentation can be found in [23]. The transformation in this method is usually performed on the gradient image rather than an intensity image.

The segmentation accuracy has been estimated by using the Dice Coefficient and it is tabulated in table 1. The output image obtained from the segmentation algorithm is compared with the ground truth image. Dice coefficient is given in equation (3),

$$\text{Dice (Image1, Image2)} = \frac{2 * tp}{2 * tp + fp + fn} \quad (3)$$

Where,

- tp – true positive value
- fp – false positive value
- fn – false negative value

Table 1. Quantitative Results of Segmentation Methods

S.No	Segmentation Methods	Accuracy
1	Binarization	82%
2	Otsu thresholding	73%
3	Watershed transform	92%

Feature Extraction

Histopathology image cell morphology plays a vital role in examining the cell structure shapes and extract meaningful image features from the segmented region. These morphological features are fed into the classifier as input data. The features like area, perimeter, eccentricity, major-axis length and minor-axis length are extracted from the nuclei segmented region and its feature values is used to distinguish the benign cells from malignant cells. The features description and its measures are shown in table 2.

Table 2. Structural Features

S.No	Structural features	Description
1	Area(A)	$A = \sum_{i=1}^n \sum_{j=1}^m S(i, j)$ Where A is the cell area and S is the segmented image of <i>i</i> rows and <i>j</i> columns
2	Perimeter(P)	$P = Evencount + \sqrt{2}(oddcoun)unit$
3	Major Axis Length (MaL)	$MaL = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ Where x_1, y_1 and x_2, y_2 are end points on the major axis.

4	Minor Axis Length (MiL)	$MiL = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ Where x_1, y_1 and x_2, y_2 are end points on the minor axis.
5	Eccentricity(E)	$E = \frac{\text{Length of major axis}}{\text{Length of minor axis}}$

For feature extraction, the output image yield from the watershed algorithm is taken as input to the feature extraction process. For each cell nuclei, structural features were estimated and it is shown in table 3.

Table 3. Structural Features Extracted from the Watershed Segmented Image for Few Cell Nuclei Samples

Cell nuclei No.	A rea	Peri meter	Major-axis length	Minor-axis length	Eccentr icity	
1.	24	2	69.98	26.357	13.945	0.84857
2.	033	8	128.1	48.4	28.279	0.81156
3.	77	5	62.50	23.11	12.764	0.83364
4.	26	4	93.82	36.98	19.698	0.84633
5.	220	5	135.9	42.743	37.138	0.49507
6.	93	6	70.96	20.94	19.992	0.29752

Classification

The features obtained from the segmented nuclei region are provided as input to the classifier. The classifiers are trained with structural descriptors like area, perimeter, eccentricity, major and minor- axis length. The range of values for each selected feature for normal and abnormal cells present in the histopathology image are tabulated in table 4. In this work, Decision tree[24], Linear Discriminant classifier[25], Support Vector Machine[26], and Random Forest classifier[27] are used to train cancerous(nuclei cell) and non-cancerous cell features. The learned model is used to predict the unknown class label of the segmented cell region in the testing stage.

Table 4. Training Features for Normal and Abnormal Cells

S .No	Structural Features	Normal breast cancer cells	Abnormal breast cancer cells
1	Area	130.05 - 386.71	457.22 - 1300.00
2	Perimeter	50.87 - 76.09	79.61 - 156.19
3	Major axis length	17.83 - 27.25	28.04 - 50.18
4	Minor axis length	10.09 - 19.44	21.47 - 40.01
5	Eccentricity	0.88 - 0.96	0.48 - 0.63

Classifier Evaluation Metrics

The performance of the classifier model is measured using various parameters such as accuracy, recall, specificity, precision and F1-score.

$$Accuracy = \frac{n_{tp} + n_{tn}}{\text{Total number of samples}} \tag{4}$$

$$Recall = \frac{n_{tp}}{n_{tp} + n_{fn}} \tag{5}$$

$$Specificity = \frac{n_{tn}}{n_{tn} + n_{fp}} \tag{6}$$

$$Precision = \frac{n_{tp}}{n_{tp} + n_{fp}} \tag{7}$$

$$F - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{8}$$

Where, n_{tp} denotes number of true positives, n_m is the number of true negatives, n_{fp} is the number of false positives and n_{fn} is the number of false negatives.

3. Results and Discussion

We used structural features for classifying cancerous and non-cancerous cells in breast histopathology images. A total of 400 histopathology slide images with cancerous and normal category images are taken for experimental purpose. The proposed framework for automatic detection of normal and abnormal cells from histopathology images consists of a pre-processing stage, region segmentation, feature extraction and classification. The original H & E image has been processed through the median filter to remove graininess noise is shown in Figure 4. The segmentation in the pre-processed histopathology image has carried out by Binarization, Otsu’s Thresholding, Watershed algorithm and their performances have been compared. The watershed algorithm outperforms other segmentation methods in terms of segmentation accuracy. In the feature extraction stage, we consider few structural descriptors as shown in table 2 are estimated from the segmented region of the cell. The features estimated are fed into SVM classifier for model construction. In our work, we used standard classifiers for validating the model to check the supremacy of the framework. We computed classifier parameters such as accuracy, misclassification rate, sensitivity, specificity and precision. The results are tabulated in table 5. In Figure 5, shows the comparative results of performance measures of various classifiers for the proposed work. The technique reported that SVM classifier performs well with 93.1% of accuracy, F-score of 94.12% and precision of 88.89%. The Random forest classifier results in a good precision value of 91.11%. The visual results of the proposed work for sample histopathology images are shown in Figure 6.

Table 5. Performance Values of Various Classifier

Classifier	Sensitivity	Specificity	Accuracy	Precision	F-score
Decision tree	88.37	76.27	83.45	84.44	86.36
Support Vector Machine	100	84.62	93.1	88.89	94.12
Linear Discriminant classifier	87.65	70.31	80	78.89	83.04
Random Forest	92.13	85.71	89.66	91.11	91.62

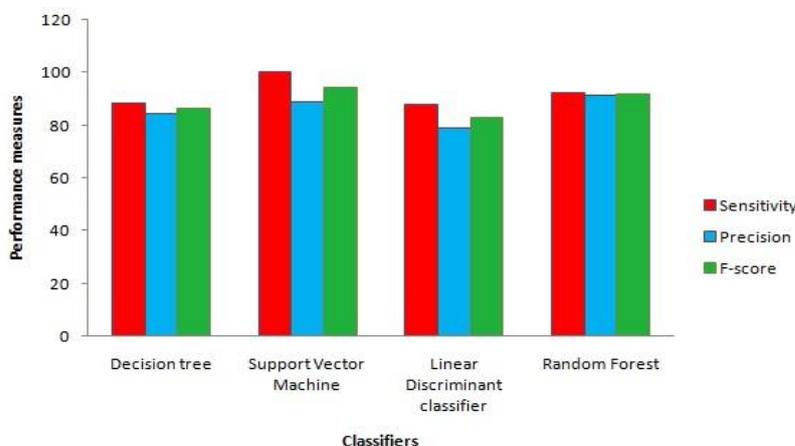


Figure 5. Comparative Results of Various Classifiers for the Proposed Work

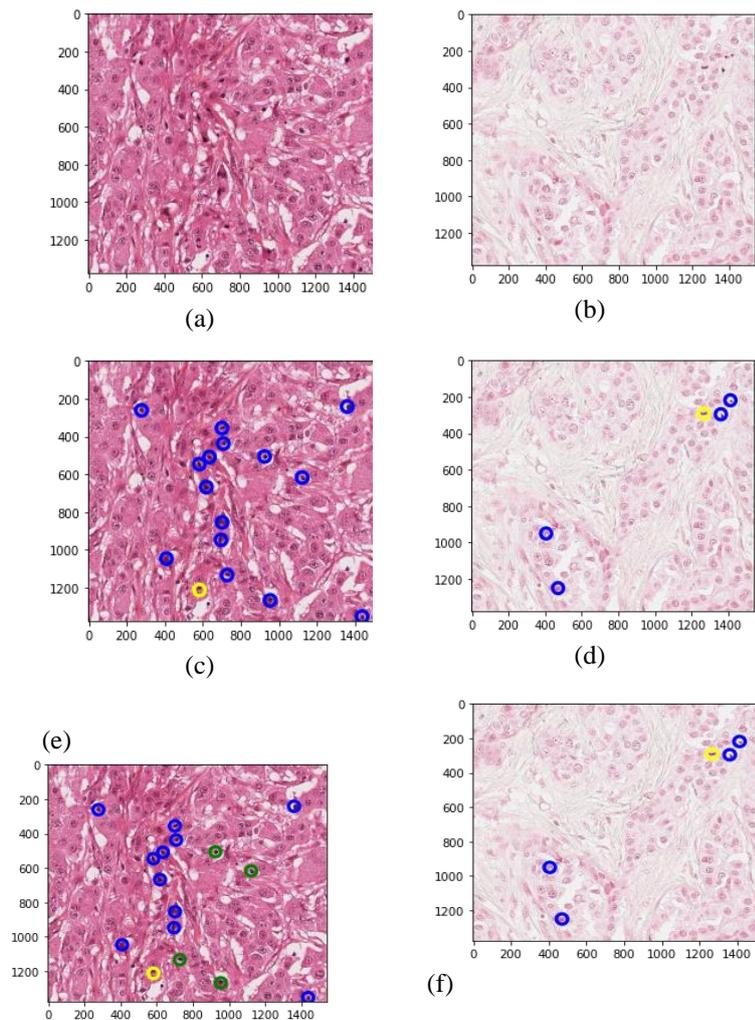


Figure 6. The Proposed Work Visual Results for Sample Images

(a) & (b) - Original histopathology images, (c) & (d) - Ground truth images,

(e) & (f) – Predicted results with cancerous cells (yellow) and non-cancerous cells (blue).

4. Conclusion

An automatic computer-aided diagnosis system helps physicians/ pathologist to achieve more efficient in cancer diagnosis. The proposed framework was used to classify the cells into cancerous and non-cancerous in histopathology breast cancer images. The approach was carried out on 400 images taken from the Mitosis Atypia dataset. Different types of nuclei segmentation approach have been used for accurate cell segmentation from background tissues. The watershed algorithm yielded a high accuracy of 82%. From the segmented region, the features are extracted and fed into the classifier for prediction. The structural features like major and minor-axis length, area, perimeter and eccentricity are determined from the segmented region of the histopathology images. Both normal and abnormal cells differ in size and shape from each other. Usually, the abnormal cells were found to be in greater magnitude. The various classifiers are built with structural features extracted from the segmentation region of 180 training images. After the prediction model constructed, the classifiers are evaluated with test dataset comprises 120 images. The accuracy of the SVM classifier is 93% and F-score of 94.12%. The learned model, Random forest classifier results in the highest precision value of 91.11% with an accuracy of 89.66% and F-score of 91.62%. To achieve more accurate detection with greater accuracy, deep learning models can be used for nuclei cell segmentation and classification.

5. Conflict of Interests

The authors declare that there is no conflict of interests regarding the publications of this paper.

References

1. Boyle P and Levin B, Eds., World Cancer Report 2008. Lyon: IRAC, (2008).
2. Siegel R L, Miller K D and Jemal A, Cancer Statistics 2018, CA: *A Cancer Journal for Clinicians*, (2018).
3. Weblink, What are the key statistics about breast cancer?
4. URL <<http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics>>
5. Ma W W and Adjei A A, "Novel agents on the horizon for cancer therapy", *CA: A Cancer Journal for Clinicians*, vol. 59, no. 2, pp. 111-137, (2009).
6. May M, "A better lens on disease: Computerized pathology slides may help doctors make faster and more accurate diagnoses", *Scientific American*, vol. 302, pp. 74-77, (2010).
7. Rubin R, Strayer D S, Rubin E, "Rubin's Pathology: Clinicopathologic Foundations of Medicine", *Lippincott Williams and Wilkins*, (2004).
8. Fischer A H, Jacobson K A, Rose J, Zeller R, "Hematoxylin and eosin staining of tissue and cell sections", *Cold Spring Harbor Protocols*, (2008).
9. Rubin R, Strayer D S, Rubin E, "Rubin's Pathology: Clinicopathologic Foundations of Medicine", *Lippincott Williams and Wilkins*, (2008).
10. Anuranjeeta, Sanjay Saxena, Shukla K K and Shiru Sharma, "Cellular Image Segmentation using Morphological Operators and Extraction of Features for Quantitative Measurement", *BioSciences Biotechnology Research Asia*, (2016).
11. Aswathy M A and Jaganath M, "Detection of breast cancer on digital histopathology images: Present status and future possibilities", *Informatics in Medicine Unlocked, Elsevier*, vol. 8, pp. 74-79, (2017).
12. Pin Wang, Xianling Hu, Yongming Li, Qianqian Liu and Xinjian Zhu, "Automatic cell nuclei segmentation and classification of breast cancer histopathology images", *Signal Processing, Elsevier*, pp.1-13, (2016).
13. Basavanahally A, Yu E, Xu J, Ganesan S, Feldman M, Tomaszewski J, Madabhushi A, "Incorporating domain knowledge for tubule detection in breast histopathology using O'Callaghan neighbourhoods", *Proceedings of SPIE*, (2011).
14. Dunder M, Badve S, Bilgin G, Raykar VC, Jain RK, Sertel O, et al. "Computerized classification of intraductal breast lesions using histopathological images", *IEEE Transactions on Biomedical Engineering*, vol.58, no.7, pp.1977-84, (2011).
15. Tosun AB and Gunduz-Demir C, "Graph run-length matrices for histopathological image segmentation", *IEEE Transactions on Medical Imaging*, vol. 30, no. 3, pp. 732-566, (2011).
16. Veta M, van Diest PJ, Kornegoor R, Huisman A, Viergever MA, Pluim JPW, "Automatic nuclei segmentation in H& E stained breast cancer histopathology images", *PloS ONE*, (2013).
17. Jain A, Atey S, Vinayak S, Srivastava V, "Cancerous cell detection using histopathological image analysis", *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, pp. 7419-7426, (2014).
18. Niwas SI, Palanisamy P, Sujathan K, "Wavelet based feature extraction method for breast cancer cytology images" *In Proceedings of the IEEE symposium on Industrial Electronics and Applications*, (2010).
19. Singh S., "Cancer cells detection and classification in biopsy image", *International Journal of Computer Applications*, vol. 38, no. 3, pp. 15-21, (2012).
20. Nie K, Chen J-H, Yu HJ, Chu Y, Nalcioglu O, Su M-Y, "Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI", *Acad radiol*, vol.15, no.12, pp.1513-1525, (2008).
21. Wilkinson M.H., Wijbenga T., De Vries G and Westenberg M A., "Blood vessel segmentation using moving-window robust automatic threshold selection," *In Proceedings of the IEEE International Conference on Image Processing*, (2003).
22. He L., Rodney L., Antani S and Thomas G.R., "Local and global Gaussian mixture models for hematoxylin and eosin stained histology image segmentation," *International Conference on Hybrid Intelligent Systems*, pp. 223-228, (2010).
23. Demir C., Yener B., "Automated cancer diagnosis based on histopathological images: a systematic survey," *Rensselaer Polytechnic Institute, Technical Report*, (2005).
24. J.B. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta Informaticae*, vol.41, no. 1, pp.187-228, (2000).
25. P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Trans. Geosci. Electron.*, vol. 15, no. 3, pp. 142-147, (1977).

26. C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467- 476, (2002).
27. S.X. Lu and X.Z. Wang, "A comparison among four SVM classification methods: LSVM, NLSVM, SSVM and NSVM," *In Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 7, pp. 4277-4282, (2004).
28. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18-22, (2002).
29. N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Transactions on System Man Cybernetics*, Vol. SMC-9, No. 1: 62-66, 1979.
30. Dataset-<https://mitos-atypia-14.grand-challenge.org>