iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VII **Month of publication:** July 2020

DOI: http://doi.org/10.22214/ijraset.2020.7073

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

Generic Imaging Heuristic for Identification of Breast Carcinoma

Dr. Tusar Kanti Mishra[1], N. Pooja Swapnika[2], M. Akhila[3], P. Aiswarya[4], U. Lahari[5]
[1, 2, 3, 4, 5]*Department of CSE, Gayatri Vidya Parishad College Of Engineering for Women, Visakhapatnam, India*

Abstract: ***Breast cancer is one of the most common cancers and it leads to death among women. Diagnosis of breast cancer manually requires more time and highly experienced pathologists. For the improvement of diagnostic consistency and efficiency, the computer assisted diagnosis systems are used in order to overcome this problem. The method proposed in this paper includes image segmentation using sliding window mechanism, extracting feature vectors by using local binary pattern algorithm and classification of histology images into benign and malignant classes using support vector machine. The approach here is applied to perform classification of breast cancer histology images and achieve certain accuracy. In this work the used algorithms are compared with various other methodologies where the proposed methodology obtain better accuracy rate and In ROC analysis, the results also contain an optimal classification performance value as the area under curve (AUC) of 0.90 which is more efficient compared to others.***

Keywords: Breast cancer histology images, sliding window, feature vectors, local binary patters, PCA, classification, SVM, ROC.

I. INTRODUCTION

To create medical images from a human body for diagnosis, the medical imaging technique is used. Different novel methods of scanning and imaging are mostly based on computer technology. The advanced medical imaging techniques, recognition to the developments in the computer science have been implemented. By using the computer technology, the medical imaging processing handles images and it consists of many different types of techniques and algorithms such as image gaining, storage, presentation, and communication. Digital image is controlled by using the image processing technique.

Now a day, one of the most dangerous health issues among women is cancer and it is increasing day by day. After lung cancer, the breast cancer is one of the most common cancers in women. Cancer is a wide word for a class of diseases characterized by abnormal cells that increase and penetrate healthy cells in the body. Group of cancer cells penetrates the periphery tissues or extend to different areas of the body leading to breast cancer. Breast cancer occurs when malignant tumors maturate in the breast. These cells are spread by parting away from the original tumor and penetrating into blood vessels or lymph vessels, which divides into tissues throughout the body. The process of damaging other tissues, organs and other parts of body is called metastasis. As its cause is not known, it is not possible to prevent this disease. However, the detection of this disease in early stage is very important to heal. Histopathology is a very efficient method for early detection of breast cancer symptoms and it identifies several abnormalities such as masses, classifications, Asymmetries, etc.

A tumor can be either benign (not harmful to health) or malignant (has the potential to be harmful). Benign tumors are non-cancerous and these cells are similar to normal tissues. They increase slowly, and do not penetrate to neighboring tissues or spread to any other parts of the body. Malignant tumors are cancerous and left without checking. Malignant cells eventually spread outside the original tumor to different parts of the body. The term “breast cancer” is defined as the development of malignant tumor in breast. The breast cancer usually develops in the cells of the lobules, which are the milk-producing glands, or the ducts, the transfers that drain milk from the lobules to the nipple. Not so commonly, breast cancer begins in the stromal tissues that include the fatty and fibrous connective tissues of the breast. Breast cancer is occurred by a genetic abnormality (a “mistake” in the genetic material).Only five to ten percent of cancers are due to an abnormal inheritance from your father or mother. Instead, eighty five-ninety percent of breast cancers are due to genetically abnormal issues. The outcome of the aging process and the “wear and tear” of the life is general. Maturation of breast cancer is due to non productive, which is either done by you or others.

Diagnosis can be done by extracting different kinds of patches which are of various sizes by a sliding window mechanism and contains cell-level and tissue-level features. Small patches that are extracted are split into multiple clusters using k-means clustering algorithm. This algorithm uses ResNet50 as a feature extractor with approximately 88.9% accuracy [1]. Instinctive extraction and classification of masses in histopology images, obtaining features and main variation between the used strategies. The main motive here is to determine or to know the benefits and drawbacks of the several methodologies.

In different with various different views which only portray and compare various methodologies significantly, this analysis also provides a significant comparison of different approaches. The production of the 7 mass detection methodologies is compared using two different breast images databases: a public digitized database. The output is shown in the form of Receiver Operating Characteristic (ROC) and free response Receiver Operating Characteristic (FROC) analysis [2].

[3] Brief information of few basic and important topics related to breast cancer identification and diagnosis are explained. This center on key CAD techniques that are novel in breast cancer diagnosis, including discovery of calcifies, obtaining the masses, obtaining architectural distortion, discovery of bilateral asymmetry, image intensification and image recuperation. [4] Introduced a process that unified data from low-level data consisting of pixel values, high-level data that depends on association among pixels, and domain-specific data that consists of data based on relationships between monograms structures for extracting the structures and segmentation of structures that are important. Morphological and nuclear features were determined to pass them to SVM classifier after using the segmentation approach and feature extraction approach.

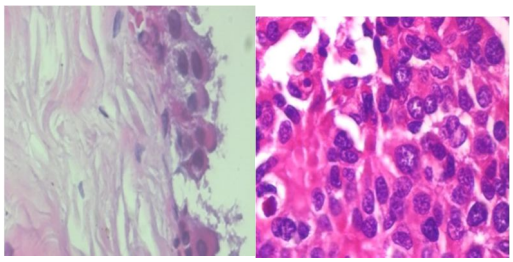
The frequently used methodology for breast cancer identification by the radiologists is Mammography. In this approach, the MIAS (Mammogram Image Analysis Society) database is used and the MIAS database comprise of cancerous and non-cancerous type of 322 breast images. The most significant step to select a quality mammogram image for further research and processing in histology image for further study and processing in histopology images analysis is pre-processing technique. Texture analysis plays efficient role to obtaining cancerous and non-cancerous types. The feature extraction process can be performed by local gradient binary patterns (LBP) operator and by using LGBP we can bethink only sign parameters, it may lose some feature information. In this approach, well known completed LBP (CLBP) methodology is used for obtaining texture feature vectors. By using the hybrid distributions of combine CLBP sign, CLBP Magnitude and CLBP Center values. LBP is one type of finished LBP for feature extraction, advantage of CLBP is rotation invariant [5]. The approach that uses the segmentation process-based on fractal texture analysis(SFTA) algorithm to extract the feature vectors from the histology images for the classification of normal ANF abnormal histology images using SVM for classification of histology images [6].

Segmentation of breast images based Fractal Texture Analysis (SFTA) to extract the feature vectors. The fractal analysis used to collect the effective information of texture features in feature vectors. Fast Correlation-based Filter (FCBF) approach used to obtain the feature subsets containing efficient features for the process of classification using SVM classifier. Texture features contains the granularity and captures the patterns and features in regions within a breast image. PCA approach optimizes the feature set [7]. The identification of biopsy tissue with H&E stain images is non-trivial and pathologists often oppose on the terminal diagnosis. To overcome these disadvantages of the feature extraction based procedures, deep learning approaches are becoming efficient alternatives. A approach for the classifier of hematoxylin and eosin stain breast histology images is by using Convolutional Neural Networks (CNNs) [8]. Using CNN the features which are extracted are used for training a SVM classifier. Four class classification has the accuracy of 77.8% and 83.3% is achieved for carcinoma/non carcinoma classification.

By referring many papers related or similar we came to know that there are many feature extraction methods for extracting features from images and many classification techniques for classification into different classes. In this paper, Image Segmentation by using sliding window mechanism, feature extraction by using local binary patterns (LBP) and principle component analysis (PCA) and binary classification using support vector machine (SVM) are proposed.

II. DATASET

The dataset consists of breast cancer histology images. Each image in the dataset is of size 700x460 pixels. These images are classified as benign (non-cancerous) and malignant (cancerous). This dataset is composed of 250 images where the trained dataset contains 200 images and tested dataset contains 50 images i.e., 80% of dataset is trained and 20% is tested. Fig. 1 shows the classified images mentioned in the dataset.



(a) (b)

FIG 1: Breast cancer histology images of each type, (a) Benign (b) Malignant

The main motive of this paper is to classify the breast cancer histopology images into benign and malignant classes with effective accuracy. The number of images present in two classes i.e. benign and malignant is shown in TABLE 1.

Table 1: The number of breast cancer histopology images per class

Class	Trained dataset	Tested dataset
Benign	100	25
Malignant	100	25
Total	200	50

III. PROPOSED METHODOLOGY

The proposed methods in this paper includes image segmentation of input image, extracting of feature vectors and binary classification. The overall process is explained in detail in this section.

A. Image Segmentation

Each image from the dataset is segmented into multiple windows by using *sliding window mechanism*. A sliding window is a rectangular region and this; mechanism "slides" the window with fixed value width and height across an image. In this algorithm, we pass the above window size through the image in the window. Then we slide the window through step size given and then convolute the next part of the image. In this way, we go cover the whole image.

The sliding window function is passed with three arguments. The first argument is the image that is going to be image. The second argument is the fixed length to move the window for sliding(step size) and the third argument is window size that is the width and height(in terms of pixels) of the window which we slide and extract from our image. The image of size 700x460 pixels is divided into multiple windows of size 70 x 46 pixels with step size 20 pixels i.e., difference between pixels of every two windows is 20 pixels. The image is taken in the gray scale format where only two dimensions of an image are considered that is width and height. Therefore, each image is divided into 672 windows by using sliding window mechanism. All these windows are given as input to the feature extraction algorithm. The fig2 describes the flow of methodology proposed in this paper.

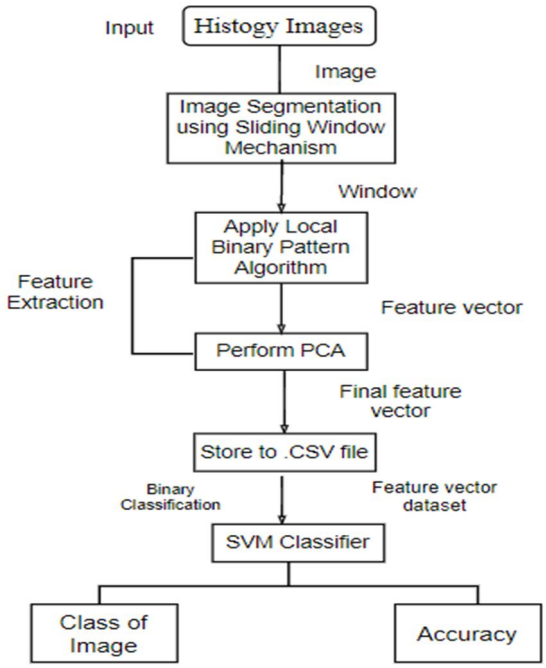


FIG 2: Block diagram for the proposed methodology.

B. Feature Extractor

The histology images have different cell morphology, texture, tissue structures, and so on. The likeness of complex features is important for the classification task. Local Binary Patterns (LBP) is one of the popular methods used in feature extraction, which was originally proposed for texture classification. Lately, it is also mostly used for other tasks such as facial expression recognition, face recognition, fingerprint identification or several medical applications such as breast cancer detection method based on LBP features. *Local Binary Pattern* (LBP) is an efficient texture operator which describes the image content, which categorize the pixels of an image by segmenting the image and values are computed by the surrounding of each pixel and considers the resultant values as a binary number. In this way, the feature vectors are created by computing the LBP histograms. To identify the cancerous tissues, classifier is used with thresholding. To provoke the real-time settings in order to analyze images, the most important computational simplicity property is used. In the grayscale image of each pixel, neighborhood size 'r' is selected by surrounding the center pixel. For this center pixel, the LBP value is computed. The input image with the same height and width is hoarded in the output 2D array. For example, let us consider an original LBP descriptor operating a fixed 3 x 3 surrounding pixels:

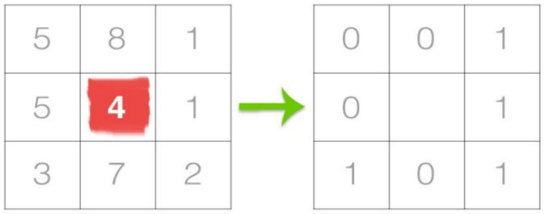


FIG 3: The initial step in LBP construction is to consider the 8 pixel from the window and compute the segmented image to get the result of set of eight binary digits.

In the above given figure we consider the center pixel (which is in red) and compute the values using the neighborhood of 8 pixels as shown in fig3. If the emphasis of the center pixel which we have segmented from the window is greater-than-or-equal-to to its surrounding we set the value to 1; otherwise we set the value to 0. Here the total possible values of LBP codes are 2^8=256 by calculating the LBP value for the center pixel. The start of pixel can be from any of the surrounding pixel and calculate our way clockwise or counter-clockwise, but the ordering of binary values must be kept constant for all pixels (in image) and images (in dataset). Considering a 3 x 3 neighborhood, so 8 neighbors are present, on which the binary set is performed and the obtained outcome is stored in an eight-bit array as described in fig4. The resulted eight-bit array is then transformed to decimal value in the following way:

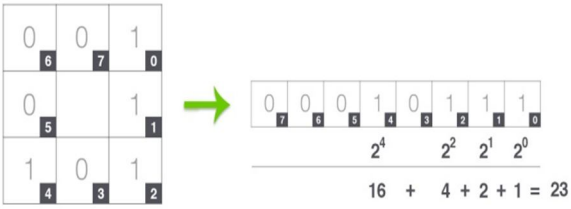


FIG 4: By considering the 8-bit binary value stored in an array is transformed into a decimal value representation.

Compute the values starting at the top-right point and calculate the values in **clockwise direction** by **acquiring** the binary string as we go along which is stored in an array as mentioned above. Then convert this binary string to decimal, resulting to the decimal value 23 as show in the fig.4. These values are stored in the 2D array which is the output of LBP. The primary use of the LBP implementation is to capture finely grained information from an image. This ability of capturing the information at such a small amount is the disadvantage of the method. Only the standard 3 x 3 scale can be captured, but varying scales cannot be captured to compute to treat variable neighborhood sizes. The solution to compute for variable neighborhood sizes, 2 parameters were defined: The number of points 'p' in a equilibrium surroundings are considered (thus remove dependence on a square neighborhood). The radius on the circle 'r', allows us to compute for different varying scales. Each histogram (from each grid) in grayscale image will contain 256 positions(0~255) representing the emphasis of each pixel and calculate the LBP values of the breast image at each and every point. The set of square cells present on a regular grid is obtained by dividing an image. For each cell as a histogram of the LBP values, the feature vector is calculated. Feature vectors are computed by using these vectors which contain p+2 features.

We compute the single value which is efficient and that represents the whole vector of a window using *Principle Component Analysis* (PCA) i.e., the Eigen values obtained for all windows are combined into single vector which is final feature vector. These feature vectors are considered as the input to the classifier.

C. Classification

Linear *Support Vector Machine* (SVM) is utilized to classify the breast cancer histology images into cancerous and non-cancerous tissue classes which take the feature vectors of images as input. To train and test the data for classification, SVM, the supervised learning method is used. Examples of the different phylum are disunited by a clear space that is as wide as possible with representation of points in space and mapped with an SVM model. The new samples are generalized in the similar area and it is estimated to belong to a level which is defined on which view they will dive into. Using kernel tricks, SVM can perform the non-linear classification efficiently and maps their inputs to feature spaces which are very high-dimensional.

For our classification we use linear SVM, where we consider a trained dataset n points which are of the model $(a_1, b_1),\ldots,(a_n , b_n)$, where bi represents each of two,1 or -1. Every point specifies the class, which these corresponds to ai . The p-dimensional real vector is obtained by each and every point of a_i. In this, the "maximal-boundary spyplane", is determined which is split into the crucial points ai for which b_i = 1,-1. This is represented as the distance between the spyplane and the adjacent point a_i in distinction to both the points which is to be increased. The spyplane refers to the rooted bit which gratify.

$$w.a - t = 0 \quad \text{- (1)}$$

Where in equation1 w = normal curve (not normalized) of the spyplane. t/||w|| describes the parameter which is the counterbalance to the spyplane from the provenance of the normal curve w. when the training data is linearly separable, 2 parallel spyplanes can be selected in which the two classes of data are separated, due to this, the separation is vast. Delimited area between two spyplanes is called as "margin". The maximal-boundary spyplane lies half way between them. By methodize or institutionalize sets of the data, the spyplanes are derives the following mathematical statement.

w.a – t = 1 (on or above the boundary of one class, with label 1) and

w.a – t = -1 (on or below the boundary of other class, with label -1)

Analytically, Separation between the two spyplanes is 2/||w||. In order to exaggerate the separation between the planes, ||w|| should be pruned. The distance among them is calculated by using the distance to a plane equation from a point. The following constraint is added in prevention of falling the due points into the margin: for each I one of two

$$w.a_i - t >= 1, \text{ if } b_i = 1, \quad \text{- (2)}$$

Or

$$w.a_i - t <= -1, \text{ if } b_i = -1, \quad \text{- (3)}$$

Each and every data points of these constraints should lie on the boundary of appropriate sidewise. By combining the equation2 and equation3 this can be rewritten as

$$b_i (w.a_i - t) >= 1, \text{ for all } 1 <= I <= n \quad \text{- (4)}$$

To get the intensification problem,it can be established together :

"Prune ||w|| exposed to $b_i (w.a_i - t) >= 1$ for i = 1,...,n".

The w and t in equation4 which solves the problem describes the classifier, a -> sign $(w.a_i - t)$. The max-margin spyplane is completely determined by that a_i which is a very important consequence of the analytical description which lies completely to those ai that lie nearest to it. The a_i mentioned in equation4 are called as support vectors.

D. Algorithm

1) Algorithm 1: Feature Extraction

Require: l: Total images

w[]: list of windows.

fv[]: feature vectors of all the images.

a) Stepsize = 20

b) (w_width, w_height) = (70 , 46)

c) for i in l do

d) w ← slidingwindow(i, w_width, w_height, stepsize)

e) for j in w do

f) f <- LBP_FeatureVector(j)

g) fv <- PCA(f)
h) end
i) end
j) return fv

2) Algorithm 2: Classification
Require: x_trained[]: feature vectors of trained images
y_trained[]: class labels of trained data
x_tested[]: feature vectors of tested images
y_tested[]: class labels of test data
w[]: width of hyper plane
b: bias value
n: number of features
a) learning_rate=0.001,lambda_param=0.01, n_iters=1000
b) w <- zeros(n), b=0
c) for _ <- 0 to n_iters do
d) for i, x_i in enumerate(x) do
e) Compute w and b
f) end
g) end
h) y_p <- predict(x_tested, w, b)
i) y_predict <- sign(y_p)
j) return y_predict

IV. EXPERIMENTAL ANALYSIS

The comparison of the performance with other histology images of breast classification is described in different methods.

A. Performance Evaluation

For evaluation and comparison of the performance of our classification of breast cancer histology images are of two categories benign (non-cancerous, non-harmful) and malignant (cancerous, harmful), the accuracy, recall, precision, confusion matrix are used as evaluation measures. The computational formulas are as follows:

Precision = TP/(TP + FP)
Recall = TP/(TP + FN)
F1 = (2*precision*recall)/ (precision + recall)
Accuracy = (TP + TN)/ (TP + FN + TN + FP)
Specificity = TN/ (TN + FP)

Here, TP (True Positive) is the count of positive cases. Respectively TN, FN and FP represent the count of True Negatives, False Negatives and False Positives. The percentage of the positive ones that are classified correctly and relevant represents the recall. To calculate the performance of multi-classification, Macro-F, which is also known as macro-averaging, is used for the calculation of the F-Scores of the n number of classes and performing average of these per-category scores for the computation of global means. The confusion matrix allows the visualization of the performance in a specific contingency table.

B. Construction and Resultants

The normalized breast cancer histology images
Explained in Section two are used to handle the analysis. The featured vectors which are derived from the histology images of breast cancer are given to SVM classifier in which the rate of learning is set to 0.001. Clustering and SVM classifier are implemented by using the machine learning algorithms with opencv and Sklearn. 90% accuracy is obtained from the classifier.

Table 2: Confusion matrix obtained by using SVM classifier.

	Predicted class	
Actual Class	Malignant	Benign
Malignant	20 (TN)	5 (FP)
Benign	0 (FN)	25 (TP)

Table 3: Performance details obtained after classification using SVM.

Class	precision	recall	F1-score	support
Benign	1.00	0.80	0.89	25
Malignant	0.83	0.90	0.91	25

At various thresholds settings, the curve of AUC-ROC measures the classification problem. ROC represents the probability and AUC represents the measure or degree of separability. It explains the capability of the model for distinguishing the classes. If the AUC curve is higher, the model is better at predicting 0s and 1s. By analogy, higher the AUC, the model is differentiates the patients with disease and without disease. The ROC curve in the graph plots with TPR and the FPR where TPR is on y-axis and FPR is on x-axis as shown in fig5.

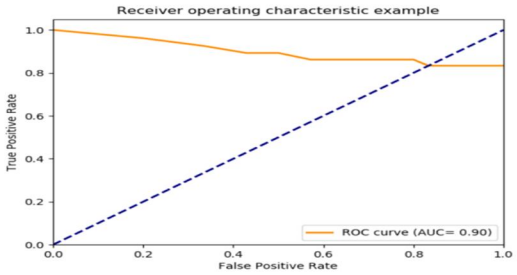


FIG 5: Plot that represent the ROC curve obtained for proposed method.

Where,
TPR/Recall/Sensitivity = TP/(TP + FN)
Specificity = TN/(TN + FP)
FPR = 1 - Specificity = FP/(TN + FP)

C. Comparative Analysis

In this paper the comparison of proposed methodologies is done with other method such as HOG, SURF, SGD, Naïve bayes algorithms as shown in table 4.As compared to other methodologies the proposed methods provide efficient accuracy with less time complexity.

Table 4: Comparison between different algorithms.

s.no.	Method	Accuracy (%)	Time Computation (sec)
1	Local Binary Patterns(LBP)	90	0.040
2	Support Vector Machine(SVM)	90	3.00
3	Histogram of oriented gradients (HOG)	49-51	1.00
4	Speeded-up robust features (SURF)	76	1.16
5	Stochastic Gradient Descent(SGD)	80	6.00
6	Naïve Bayes	50	16.00

The comparison between the accuracy rate of LBP and SURF algorithms for different sizes of test data is shown in the fig6.

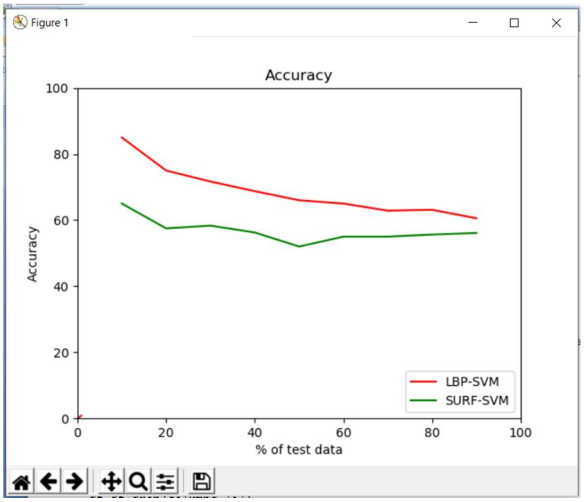


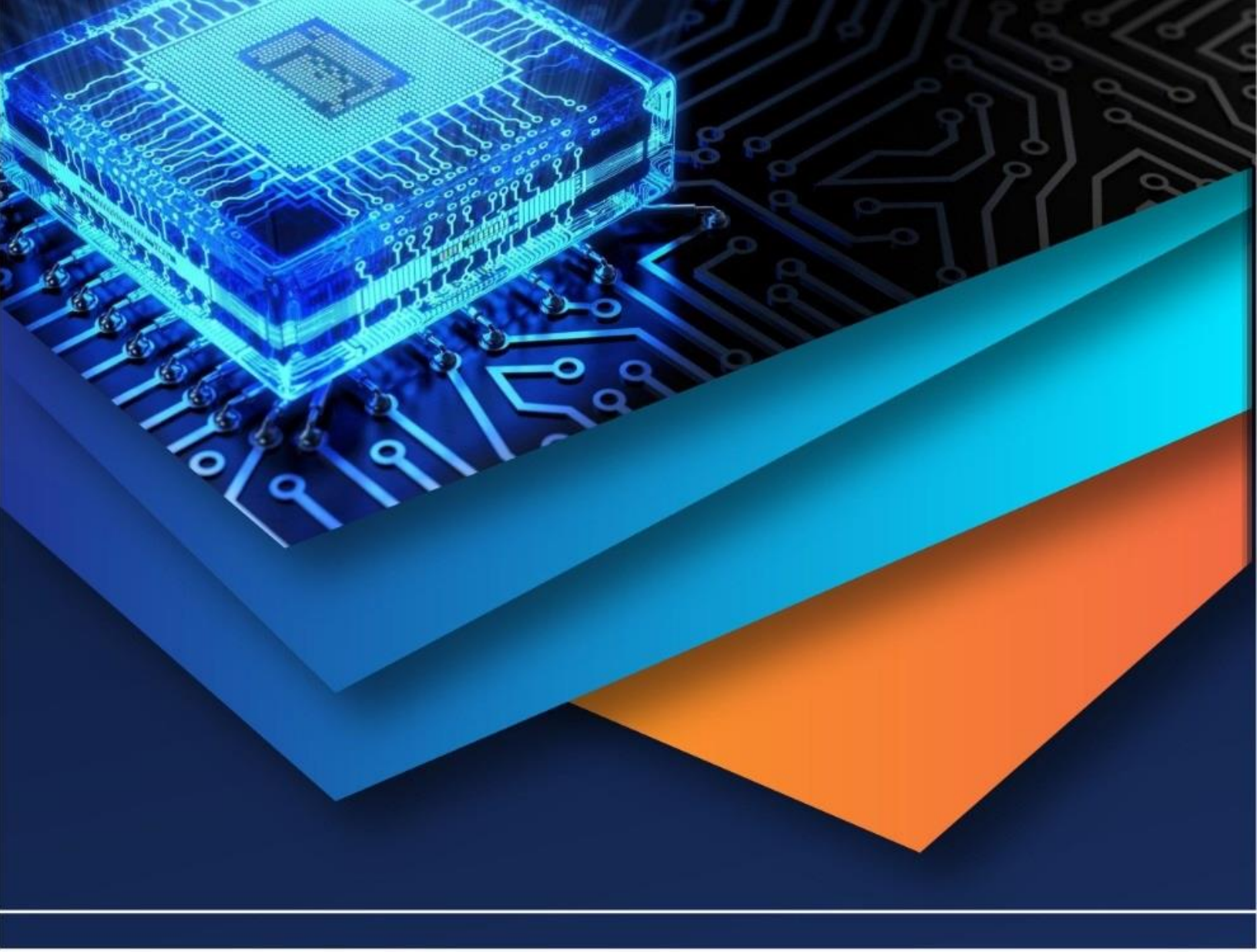
FIG 6: Plot that shows the comparison of accuracy rate of LBP and SURF methods.

V. CONCLUSION

In this formal, an efficient methodology is proposed to classify the stained breast cancer histopology images into two classes: Benign (cancerous) and Malignant (non-cancerous). The process is designed in which the image segmentation is included using sliding window mechanism, feature extraction using LBP, PCA and classification by using SVM classifier. This achieves 90% accuracy.

REFERENCES

[1] Yuqian Li, Junmin Wu and Qisong Wu, "Classification of Breast Cancer Histology Images using Multi-Size and Discriminative Patches Based on Deep Learning", Special Section on Deep Learning for Computer-Aided Medical Diagnosis, feb7, 2019.
[2] Arnau Oliver, Jordi Freixenet, Joan Marti, Elsa Perez, Josep Pont, Erika RE Denton, and Reyer Zwiggelaar, "A Review of Automatic Mass Detection and Segmentation in Mammographic Images", Medical Image Analysis, vol. 14, no. 2, pp. 87-110, 2010.
[3] Jinshan Tang, Rangaraj M Rangayyan, Jun Xu, Issam Ei Naqa, and Yongyi Yang, "Computer-Aided Detection of Diagnosis of Breast Cancer with Mammography: Recent Advances", Information Technology in Biomedicine, IEEE Transactions, vol. 13, no. 2, pp. 236-251, 2009.
[4] S.Naik, S.Doyle, A.Madabhushi, M.Feldman, J.Tomaszewski, and S.Agner, "Automated Gland and Nuclei Segmentation for Grading of Prostate and Breast Cancer Histopathology", in Proc. 5th IEEE Int, Symp., May 2008, pp.284-287.
[5] S Naresh and S Vani Kumari, "Breast Cancer Detection using Local Binary Patterns", International Journal of Computer Applications, vol. 123, no. 16, 2015.
[6] A. Dong and B.Wang, "Feature Selection and Analysis on Mammogram Classification", In Communications, Computer and Signal Processing, 2009. PacRim2009. IEEE Pacific Rim Conference on, pages731-735. IEEE, 2009.
[7] Shradhananda Beura, Banshidhar Majhi, Ratnakar Dash, "Automatic Characterization of Mammograms using Fractal Texture Analysis and Fast Correlation based Filter Method", Association for Computing Machinery, feb26-27, 2015.
[8] T.Araujo et al., "Classification of Breast Cancer Histology Images using Convolutional Neural Networks", PLoS ONE, vol. 12, no. 6, p. e0177544, 2017.
[9] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam, "Local Binary Patterns Variants as Texture Descriptors for Medical Image Analysis", Artificial Intelligence in Medicine, vol. 49, no. 2, AP, 117-125, 2010.
[10] N.Chritian. (2018).What You Need to Know About Breast Cancer. [Online]. Available: https://www.medicalnewstoday.com/articles/37136.php
[11] A.R.Backesand O.M.Bruno. "A New Approach to Estimate Fractal Dimension of Texture Images", In Image and Signal Processing, pages 136-143. Springer, 2008.
[12] "Breast Cancer detection using image processing techniques", Tobias Christian, Melanie.A, James C. IEEE International Conference. February 2000.
[13] American Cancer Society(1999).Breast Cancer Facts & Figures 1997-1998.
[14] Basset L.W., Manijikian V. III, Gold R.H. (1990). Mammography and breast cancer screening (Review). Surg Clinics of North America, 775-800.
[15] Kopans, D.B. (1996). The potential impact of computer-aided diagnosis on clinical mammography. Proceedings of the Third International Workshop on Digital Mammography, 35.
[16] "Computer-Aided Detection and Diagnosis in Momography". Mehul S, Mia K.M. Handbook of Image and Video Processing. Edition: 2nd. January 2005.
[17] E. J. Feuer, L. Wun, C. C. Boring, W. D. Flanders, M. J. Timmel, and T. Tong, "The Lifetime Risk of Developing Breast Cancer," Journal of the National Cancer Institute, vol. 85, pp. 892-897, 1993.
[18] M. L. Giger, "Computer-aided diagnosis of breast lesions in medical images," Computing in Science & Engineering, vol. 2, pp. 39-45, 2000.
[19] "Support Vector Machines:Theory and Applications". Theodoros E, Massimiliano P. Machine Learning and Its Applications, Advanced Lectures.
[20] "Data Classification Using Support Vector Machine". Durgesh K.S., Lekha B. Journal of Theoritical and Applied Information Technology.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY