IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 8 (August. 2019), ||S (I) || PP 44-52

Classification of Benign and Malignant Breast Cancer Cells

Saibee Alam[1], Kirti Gaur[2]
Department of Computer Science and Engineering, Rajasthan Technical University, Kota, Rajasthan

Abstract: Breast Cancer is a malignant Tumor (a collection of cancer cells) that arises from the cells of breast. In all over the world it is a major health issue. In this paper a model is described for the classification of malignant and benign breast cancer cells. In this model Gaussian filter is used for noise reduction and smoothening. The Gaussian filter is applied on two parameters of a feature. After the filtering, the normalized values of the parameters are produced. The produced result is classified using SVM classifier. The proposed method uses SVM classifier with the combination of Gaussian filter.

Keywords: *Breast Cancer, Gaussian Filter, Histopathological, Support Vector Machine.*

Date of Submission: 06-08-2019 Date of Acceptance: 22-08-2019

I. INTRODUCTION

Breast Cancer is a malignant tumor (a collection of cancer cells) that arises from the cells of breast. Not only in western country but also in our country, it is a major health problem. It is the 2nd most common cancer which is diagnosed world-wide (Parkin et al. 2001). Breast cancer usually occurs in the ducts, pipes that convey milk and organ that produces milk. It occurs both in men and women but in men it is an uncommon disease. We can determine the maturity of a cell by examining its parameters, such as radius, texture, concavity etc. Numerous masses are benign whereas the malignant tumor cells are extremely serious cancer. Before attaining a tangible size, the larger part of breast tumor will have metastasized. On the basis of the parameters of the tumor cell we classified it into benign or malignant class.

The existing technique for Breast Cancer classification using Random Forest [2] is not giving the accurate results as its accuracy is 86%. To get more accurate results we need to make improvement. We described a model for classification of benign and malignant breast cancer cell. The Gaussian filter is used for noise reduction and smoothening. By applying the Gaussian filter on two parameters of a feature the normalized values are produced. The produced result can classify using SVM classifier. The proposed method gives the accuracy of 97.5% using SVM classifier with the combination of Gaussian filter.

Breast cancer histopathological dataset (.csv) is used. In this features are calculated from a digital image of a fine needle aspirate (FNA) of a breast lump. It describes characteristics of the cell nuclei present in the image. It has the data of 570 images, having 2 classes- Benign and Malignant. In the benign tumor cells are not cancerous and would not spread. Whereas in the malignant tumor Cells are highly cancerous and spread to other tissues and organs. In this dataset for each cell nucleus, ten real-valued features Calculate:[21]

a) Radius is average distance from center point to perimeter point.
b) Texture is Standard deviation σ of values of gray-scale.
c) Perimeter
d) Area
e) Smoothness is defined as local variation in radius lengths.
f) Compactness = [(perimeter)2 / area] - 1
g) Concavity is the severity of the contour recess.
h) Concave points are the number of concave parts of the outline.
i) Symmetry
j) Fractal dimension= "coastline approx." – 1

These features are divided into three parts first is Mean (3-13), Stranded Error (13-23) and Worst (23-32) and each contain above 10 parameter.

We applied Gaussian filtering on the “radius” feature. Applying filter on a particular feature will have effect on the values of its correlated feature. To overcome this effect on the values of correlated features, we discarded the features which are correlated to the “radius”. The “radius” and the “smoothness” both are correlated so we discarded “smoothness” from the dataset. As the result we have new dataset with 9-features in total. Each divided into 3-parts Mean, Stranded Error and Worst.

II. METHODOLOGY

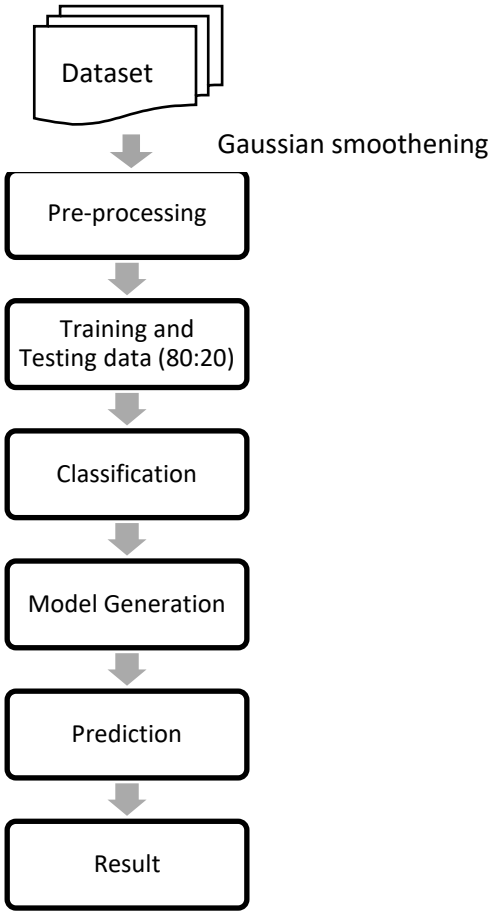

Figure 1: Proposed Method

A Breast Cancer histopathological dataset is used for classification. Gaussian smoothening is applied for noise reduction on the dataset as the pre-processing then the dataset is divided into 80:20 training and testing dataset using 5-fold cross validation. Classification is done using support vector machine as it gives the better results. The generated model is used for prediction of malignant and benign breast cancer cell.

2.1 Gaussian Smoothening of the Dataset

Gaussian filter is used to reduce the noise. It is same as convolving the image with a Gaussian function. It reduces the image's high-frequency components. It is an image blurring filter that uses a Gaussian function to calculate the transform which is applied to each pixel in the image. In 1D, the Gaussian function is:[14]

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

Where σ is the standard deviation of the distribution. It is supposed to be the mean of the distribution is 0. In 2D, the Gaussian function is:[14]

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{x^2+y^2}{2\sigma^2}} \quad (2)$$

The Gaussian filter is a non-uniform low pass filter. The larger value of standard deviation produces a wider peak that means greater blurring.The kernel size of the Gaussian function must increase with increasing σ to maintain the Gaussian nature of the filter. Gaussian kernel coefficient depends on the value of σ. At the edge of the mask, coefficient must be close to 0 [14]. Gaussian kernel is separable, which allows fast computation.it might not preserve image brightness. It is separable because we can write the 2D Gaussian function into the form of 1D Gaussian function, in other words the 2D function can be written as the combination of two 1D functions (one for X-axis and another for Y-axis).

The Gaussian filter is applied on 2 parameters of a feature to reduce the noise. Here the standard deviation σ = 10. The two parameters are radius_mean and radius_worst. We applied the filter only on these two parameters because the variation in the value of these parameters was very high so the filtering normalizes the values.
Fig. 2.shows the 1D Gaussian distribution having standard deviation equal to 10 (from 10-100) and mean equal to 50. As we can see the plot is having the wider peak that means it will give the greater blurring of the image.

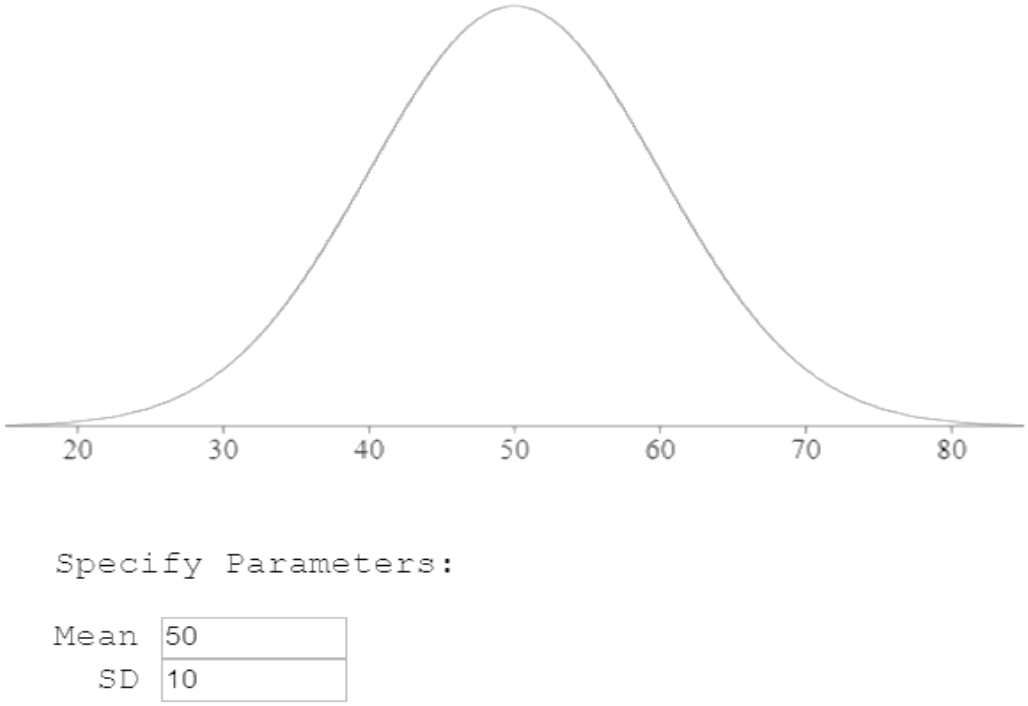

Figure 2: Gaussian distribution having σ = 10 and μ= 50

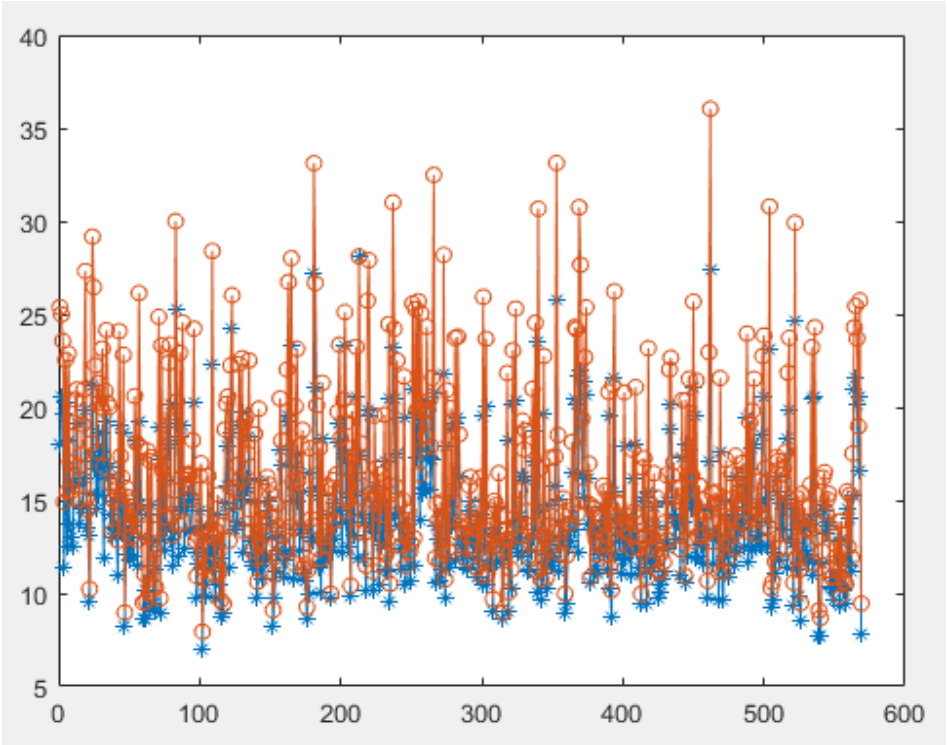

Figure 3:radius_mean and radius_worst before filtering

Fig. 3.shows the plot of radius_mean and radius_worst before filtering. The plot gives the relationship between the image and its two parameters. In this plots X-axis shows the row value (image no.) and Y-axis shows the

column values (the measure of parameters radius_mean and radius_worst of the corresponding image). Here "red o" denotes radius_worst and "blue *" denotes radius_mean.
We are taking 266th image as an example for later comparison The value of radius_worst corresponding to 266th image before filtering is 32.49 µm. and the value of radius_mean corresponding to 266th image before filtering is 20.73 µm. We can see the variations in the values of these two parameters in the plot.
Fig. 4.gives the relationship between radius_worst and filtered radius_worst. In this plot X-axis shows the image no. and Y-axis shows the the measure of parameters radius_worst and filtered radius_worst of the corresponding image.

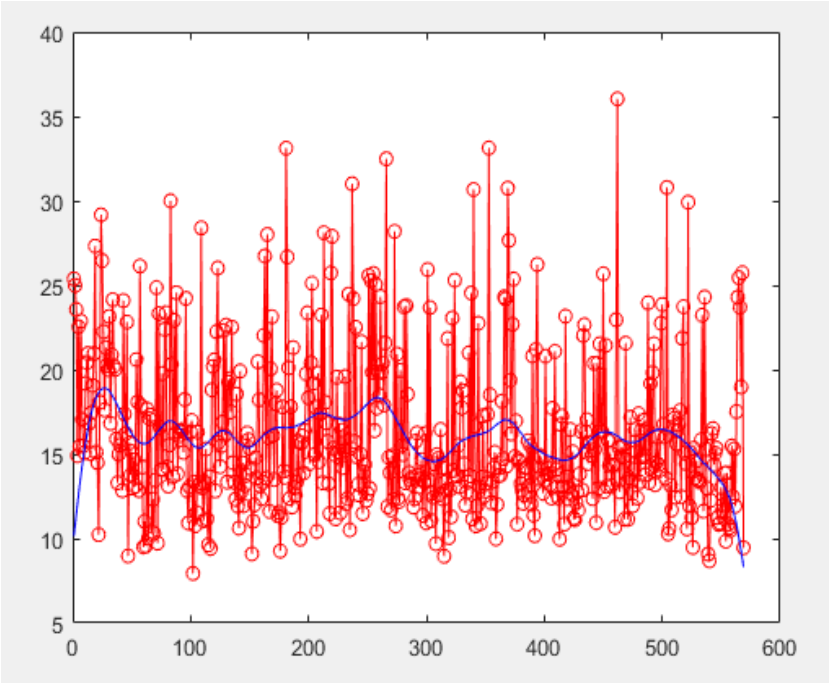

Figure 4:radius_worst and filtered radius_worst

Here "red o" denotes radius_worst before filtering and the solid blue line shows the values of filtered radius_worst. The value of radius_worst corresponding to 266th image before filtering is 32.49 µm. and the value of radius_worst corresponding to 266th image after filtering is 18.0129 µm. We can see the difference in both values. In the above plot the variation in the values of radius_worst ("red o") are very high but after filtering we can see that the variation in the values of filtered radius_worst (solid blue line) are very low.

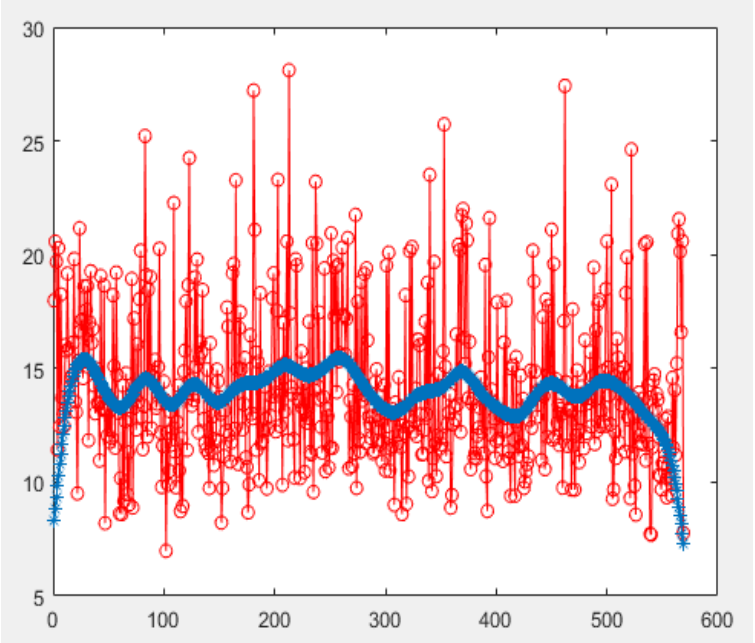

Figure 5:radius_mean and filtered radius_mean

Fig. 5.shows the plot of radius_mean before and after filtering. In these plots X-axis shows the row value (image no.) and Y-axis shows the column values (the measure of parameters radius_mean and filtered radius_mean of the corresponding image). Here "red o" denotes radius_mean before filtering and the "blue *" shows the values of filtered radius_mean. The value of radius_mean corresponding to 266th image before filtering is 20.73 µm. The value of radius_mean corresponding to 266th image after filtering is 15.2599 µm. We can see the difference in both values. In the above plot the variation in the values of radius_mean ("red o") are very high but after filtering we can see that the variation in the values of filtered radius_mean (blue "*") are very low.

Fig. 6.gives the relationship between the image and its parameter. In this plot X-axis shows the row value (image no.) and Y-axis shows the column values (the measure of parameters filtered radius_worst and filtered radius_mean of the corresponding image). Here "red o" denotes radius_worst after filtering and the "blue *" shows the values of radius_mean after filtering. Fig. 6.shows the plot of both parameters filtered radius_mean and filtered radius_worst. The value of filtered radius_mean and filtered radius_worst corresponding to 266th image is 15.2599 µm and 18.0129 µm respectively. We can compare the Fig. 3. and Fig. 6. to see the difference in values after filtering.

Table 1 shows the comparison of filtered and non-filtered values of 266th image.

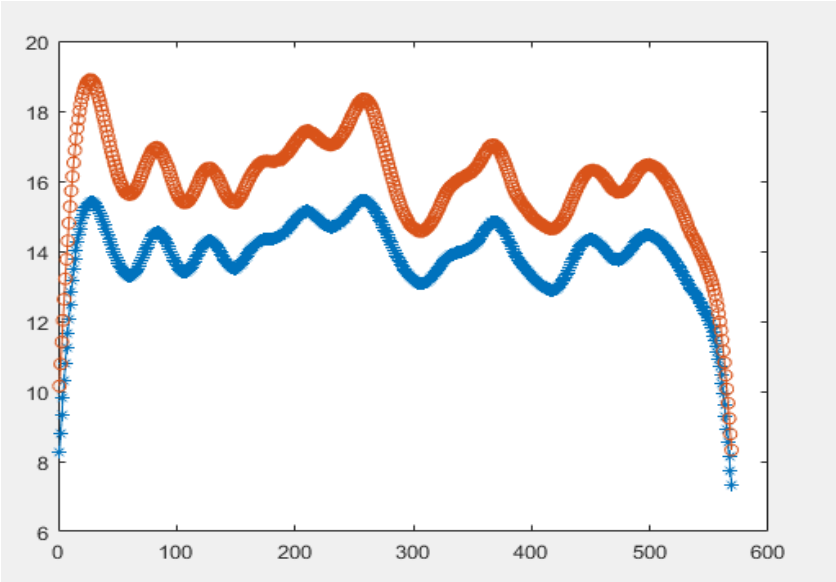

Figure 6: Filtered radius_worst and filtered radius_mean

Table 1: Comparison of Parameters Before and After Filtering

Parameters	**Values (266th image) in µm**
radius_mean	20.73
Filtered radius_mean	15.2599
radius_worst	32.49
Filtered radius_worst	18.0129

2.2 Classification of Filtered Dataset

The SVMs are set of related supervised learning methods used for classification and regression [19]. They are member of a family of generalized linear classification [20]. The actual classification error is minimized and the geometric margin is maximized by SVM, because of this special property SVM known as maximum margin classifier [20]. It is dependent on the mechanism called as structural risk minimization (SRM). An input vector is mapped to a higher dimensional space using SVM, where a maximal separating hyperplane is formed.

To separate the data, two hyperplanes are formed parallely at each side of the hyperplane. The hyperplane that enlarge the distance between the two parallel hyperplanes is called as separating hyperplane. A relation is defined between the 'distance between parallel hyperplane' and 'generalization error', that larger distance will give better generalization error of the classification. The data points are considered in the following form:

$$\{(a_1,b_1),(a_2,b_2),(a_3,b_3)\ldots\ldots\ldots(a_n,b_n)\}$$

Where b_n = 1 / -1, A class (denoted by a constant) from which an belongs, n= number of samples, an = p-dimensional real vector.

To guard against attributes with length variance the scaling is important. The training data can be viewed by separating the hyperplane, which takes:

$$wa+s=0 \qquad (3)$$

Where, s= scalar vector, w= p-dimensional vector.

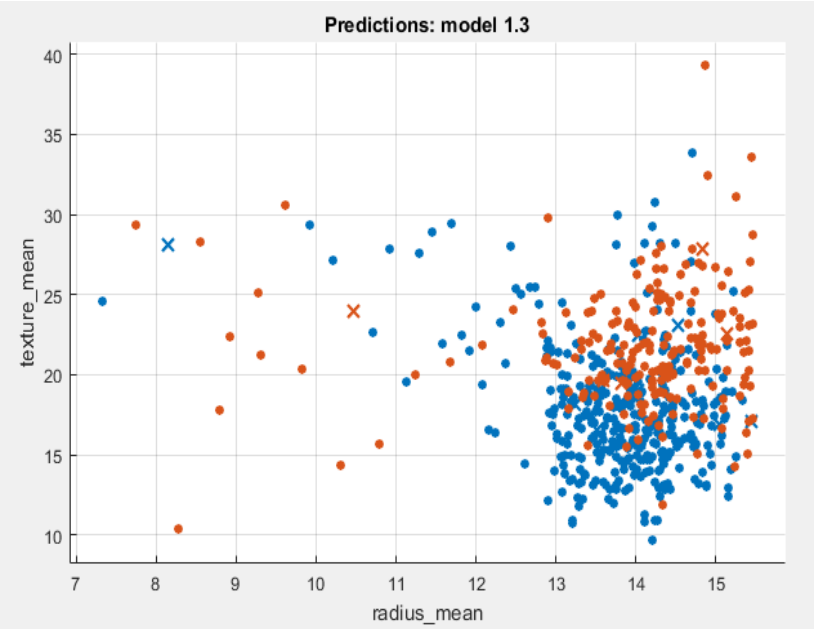

Figure 7: Scatter plot

Fig. 7.Shows the scatter plot of model. This scatter plot is plotted between the parameters radius_mean and texture_mean. The X-axis denotes radius_mean and the Y-axis denotes texture_mean. Here the blue dot refers right prediction of benign cell, red dot shows right prediction of malignant cell. Blue Cross shows the wrong prediction which has true class as malignant and predicted class as benign. The Red Cross shows the wrong prediction which has true class as benign and predicted class as malignant.

Confusion matrix is used to calculate classification accuracy of our classification model. It describes the performance of our classification model. Fig. 8.shows the confusion matrix between predicted class and true class. Here 352 benign cells are predicted correctly (TP). 203 malignant cells are predicted correctly (TN). 9 malignant cells are predicted as benign and 5 benign cell is predicted as malignant.

Classification accuracy = $[(TN+TP)/(TN+TP+FN+FP)] \times 100$ (4)

Error rate = $[(FN+FP)/(TN+TP+FN+FP)] \times 100$ (5)

True positive rate (TPR) = $TP/(TP+FN)$ (6)

True negative rate (TNR) = $TN/(TN+FP)$ (7)

False positive rate (FPR) = $FP/(TN+FP)$ (8)

Using (4), we will calculate overall classification accuracy of our model and (5) is used to calculate error rate or misclassification rate. Putting TP = 352, TN = 203, FP = 9, FN = 5 we get following results:

Classification accuracy = 97.5%

Error rate = 2.46%

True positive rate (TPR) ~ 1

True negative rate (TNR) = 0.957

False positive rate (FPR) = 0.042

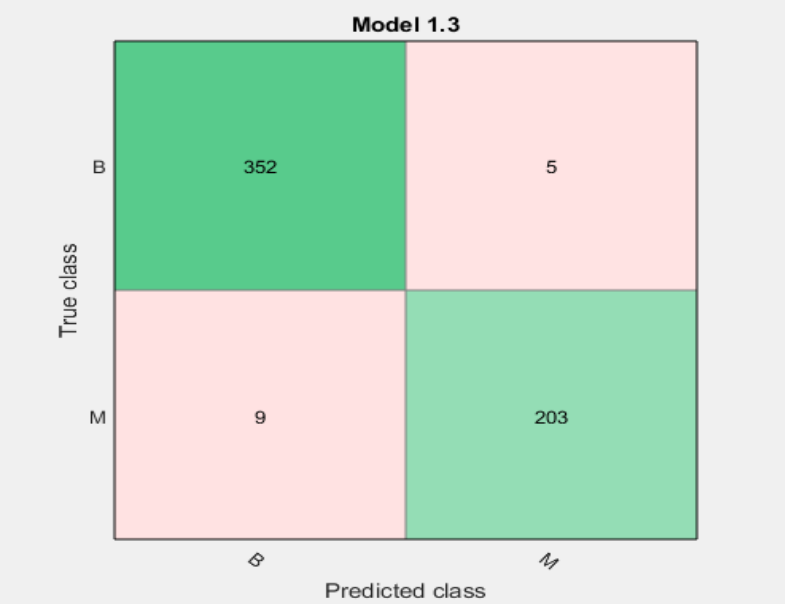

Figure 8: Confusion Matrix

Reverse Operating Characteristic curve is common way to visualize the performance of a "binary classifier". We use the ROC curve which is used to evaluate the performance of a classifier, and produce a higher score for the same classifier in comparison to other suitable classifiers, that is the main objective of AUC (Area under the Curve). AUC is exactly the percentage of the blue box that is under the curve as shown in Fig. 9. SVM classifier has an AUC of close to 1, which means classifier is perfect.

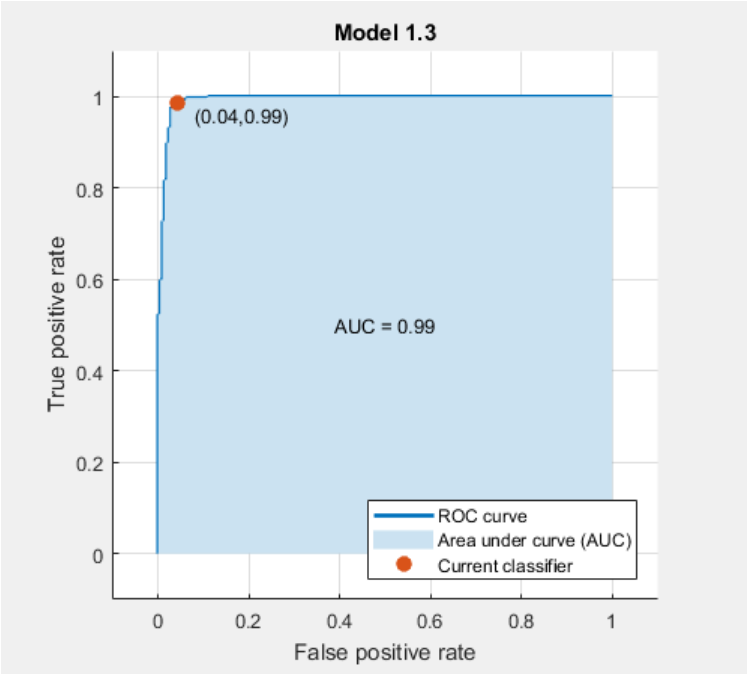

Figure 9:ROC Curve

III. RESULT

After implementing the proposed method we got following results:

Table 2: Results

Filter	**Classifier**	**% Accuracy**
Gaussian Filter	**Cubic SVM**	**97.50**
	Medium Gaussian SVM	96.80
	RF	76
	Naive Bayes	93.60
	KNN	95

Table 3: Comparison with different approaches

Filter	**Classifier**	**%Accuracy**
DTG filter + LBP [2]	RF	84
LBP [3]	1NN	75.60
	RF	74
	SVM	74.20
Gaussian filter	**Cubic SVM**	**97.50**

IV. CONCLUSION

In the proposed method, the breast cancer stroma is classified in two classes benign and malignant cells using Gaussian filter combined with support vector machine. The classification accuracy of this method is 97.5%.

The proposed method can be applied to detect other types of cancer such as brain tumor. The future work includes the same detection for more multiple classes.

REFERENCES

[1]. Yuqian Li, Junmin Wu, Qisong Wu, "Classification of Breast Cancer Histology Images Using Multi-Size and Discriminative Patches Based on Deep Learning". *IEEE Access*, Vol. 7, pages 21400 – 21408, 2019.

[2]. Sara Reis, Patrycja Gazinska, John H. Hipwell, "Automated Classification of Breast Cancer Stroma Maturity from Histological Images" *IEEE Transactions on Biomedical Engineering* DOI 10.1109/TBME.2017.2665602

[3]. Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, Laurent Heutte "A Dataset for Breast Cancer Histopathological Image Classification". *IEEE Transactions on Biomedical Engineering*, vol. 64, 2017.

[4]. Tooba Salahuddin, Fatima Haouari, Fahad Islam, Rahma Ali, Sara Al-Rasbi, Nada Aboueata, Eman Rezk, Ali Jaoua, "Breast cancer image classification using pattern-based Hyper Conceptual Sampling method", *Elsevier Informatics in Medicine Unlocked*, Vol. 13, Pages 176-185, 10 July 2018.

[5]. Xingyu Li, Marko Radulovic, Ksenija Kanjer, Konstantinos N. Plataniotis, "Discriminative Pattern Mining for Breast Cancer Histopathology Image Classification via Fully Convolutional Autoencoder", *IEEE Access*, vol. 7, pages 36433 – 36445, 2019.

[6]. Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, Anant Madabhushi, "Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images" *IEEE Transactions on Medical Imaging*, vol. 35, pages 119-130, 2016.

[7]. Sami S. Brandt, Gopal Karemore, Nico Karssemeijer, Mads Nielsen, "An Anatomically Oriented Breast Coordinate System for Mammogram Analysis", *IEEE Transactions on Medical Imaging*, vol. 30, Pages 1841 – 1851, 2011.

[8]. M. Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, Metin N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images", *IEEE Transactions on Biomedical Engineering*, vol. 58, pages 1977 – 1984, 2011.

[9]. Po-Hsiang Tsui, Yin-Yin Liao, Chien-Cheng Chang, Wen-Hung Kuo, King-Jen Chang, Chih-Kuang Yeh, "Classification of Benign and Malignant Breast Tumors by 2-D Analysis Based on Contour Description and Scatterer Characterization", *IEEE Transactions on Medical Imaging*, vol. 29, pages 513-522, 2010.

[10]. Mehmet C. Kale, Bradley D. Clymer, Regina M. Koch, Johannes T. Heverhagen, Steffen Sammet, Robert Stevens, Michael V. Knopp, "Multispectral Co-Occurrence With Three Random Variables in Dynamic Contrast Enhanced Magnetic Resonance Imaging of Breast Cancer", *IEEE Transactions on Medical Imaging*, vol. 27, pages 1425-1431, 2008.

[11]. Douglas J. Kurrant, Elise C. Fear, David T. Westwick, "Tumor Response Estimation in Radar-Based

Microwave Breast Cancer Detection", *IEEE Transactions on Biomedical Engineering*, vol. 55, pages 2801-2811, 2008.
[12]. Anna N. Karahaliou, Ioannis S. Boniatis, Spyros G. Skiadopoulos, Filippos N. Sakellaropoulos, Nikolaos S. Arikidis, Eleni A. Likaki, George S. Panayiotakis, Lena I. Costaridou, "Breast Cancer Diagnosis: Analyzing Texture of Tissue Surrounding Microcalcifications", *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, pages 731-738, 2008.
[13]. Jacob Levman, Tony Leung, Petrina Causer, Don Plewes, Anne L. Martel, "Classification of Dynamic Contrast-Enhanced Magnetic Resonance Breast Lesions by Support Vector Machines", *IEEE Transactions on Medical Imaging*, vol. 27, pages 688-696, 2008.
[14]. https://www.cs.auckland.ac.nz/courses/compsci373s1c/PatricesLectures/Gaussian%20Filtering_1up
[15]. Mithelesh Kumar Kaushik, Mrs. Rashmi Kashyap, "A Review Paper on Denosing Filter using 2D Gaussian Smooth Filter for Multimedia Application", *International Journal of Advanced Research in Computer Science and Software Engineering* vol. 3, pages 21-26, 2016.
[16]. Ruchika Chandel, Gaurav Gupta, "Image Filtering Algorihtms and Techniques: A Review", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, pages 198-202, 2013.
[17]. Pooja kamvisdar, Sonu Saluja, Sonu Agarwal, "A survey on Image Classification Approach and Techniques", *International Jurnal of Advance Research in Computer and Communication Engineering* vol.2, isssue 1 jan,2013.
[18]. Durgesh K. Srivastava, Lekha Bhambhu, "Data classification using support vector machine", *Journal of Theoretical and Applied Information Technology*, 2009.
[19]. C.-W. Hsu and C. J. Lin. "A comparison of methods for multi-class support vector machines", *IEEE Transactions on Neural Networks*, 13(2):415-425, 2002.
[20]. Durgesh K. Srivastava, Lekha Bhambhu, "Data classification using support vector machine", *Journal of Theoretical and Applied Information Technology*, 2009.
[21]. Emad A. El-Sebakhy, Kanaan Abed Faisal, T. Helmy, F. Azzedin, and A. Al-Suhaim "Evaluation of Breast Cancer Tumor Classification with Unconstrained Functional Networks Classifier", *IEEE International Conference on Computer Systems and Applications*, 2006.

Saibee Alam. "Classification of Benign and Malignant Breast Cancer Cells." IOSR Journal of Engineering (IOSRJEN), vol. 09, no. 08, 2019, pp. 44-52.