Breast Cancer Detection Using Image Processing

S. Haripriya[1], M. Havena[2], M. Rakshanabegam[3], B. Rosemaa Devi[4], P.Selvan[5]
Department of Electronics and Communication Engineering,
Chettinad College of Engineering and Technology, Karur – 639 114
Email: harislk2000@gmail.com, mhavena29@gmail.com, mrakshnabegam042000@gmail.com, rosemaadevi@gmail.com, selvanece45@gmail.com

***Abstract*— Breast cancer is one of the second leading causes of cancer death in women. Despite the fact that cancer is preventable and curable in primary stages, a huge number of patients are diagnosed with cancer very late. Conventional methods of detecting and diagnosing cancer mainly depend on skilled physicians, with the help of medical imaging, to detect certain symptoms that usually appear in the later stages of cancer. Therefore, we present here the computerized method for cancer detection in its early stage within a very short time. Here, we have used Machine learning to train a model using the predicted features of the nuclei of cells. A comparative study of two different algorithms KNN and SVM is conducted where the accuracy of each classifier is measured. After this, we analyze a digital image of a fine needle aspirate (FNA) of breast tissue using image processing to find out the features of nuclei of the cells. We then apply the feature values to our trained model to find whether the tumor developed is benign or malignant.**

***Index Terms*— Machine learning • Fine needle aspirate (FNA) • Image processing, KNN, SVM, Benign, Malignant.**

I. INTRODUCTION

Cancer is one of the main causes of human death and a major public health concern worldwide. Cancer refers to the uncontrolled growth and propagation of cells. It appears in almost any part of the body when a cell accumulates a set of mutations, generally during various years. Growth promoting genes in normal cells are duplicated several times in cancer cells and often become unstable acquiring lethal characteristics as they multiply. The American Cancer Society estimates about 1,735,350 new cancer cases diagnosed and 609,640 cancer deaths in the US by 2018.

Breast cancer is an uncontrolled growth of cells that starts in the breast tissues and affects both women and, very rarely, men (less than 1% of all breast cancer cases). In its initial stages, breast cancer can be detected by means of image studies (mammography, ultrasound, and magnetic
resonance imaging) or, less often, clinical trials of palpable tumors. Recently, strategies to mitigate breast cancer have focused on prevention and its early detection and treatment. To detect the breast cancer, most available tests are mammography and tom synthesis. Here we using the new technique with KNN algorithm and SVM algorithm.

II. BREAST CANCER

Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells

Grenze ID: 02.ICIEICE.2021.1.26

called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can grow and spread to other parts of the body. A benign tumor means the tumor can grow but will not spread. The most kinds of breast cancers are Invasive ductal carcinoma, the cancer cells grow outside the ducts into other parts of the breast tissue, Invasive cancer cells can also spread, or metastasize, to other parts of the body and invasive lobular carcinoma, Cancer cells spread from the lobules to the breast tissues that are close by. These invasive cancer cells can also spread to other parts of the body.

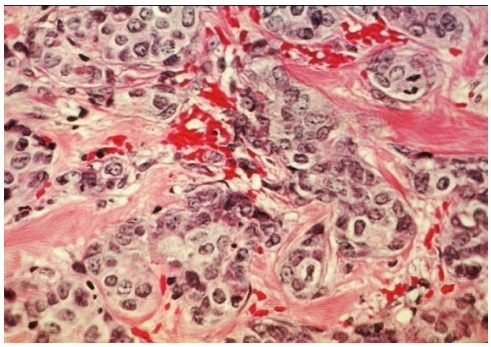


Figure 1: breast cancer cells

III. Existing work Vs. Proposed work

We have already studied two papers; in this paper they explain detection of breast cancer using mammography and tom synthesis. But the accuracy level is very low to detect the breast cancer. To rectify the drawbacks in the previous methodologies, we are using an image processing by Matlab software. We
have tried to overcome the limitations of diagnosis by proposing algorithms for finding out the features and have applied various machine learning algorithms to predict the malignancy. First section of the methodology includes the various techniques that we have proposed for the calculation of features, and the second part includes the techniques of classification that are used for prediction. In the first part using various image processing techniques, we have tried to build our own algorithm by using these concepts. The proposed methodology for processing the images of FNA slides is discussed in detail in the upcoming sections. The features that are in the breast cancer prediction data set are calculated by using this method.

IV. Methodology

First, the steps for tracing the contours of the nuclei are discussed for extracting the feature values. The steps for contour tracing are in fig.2

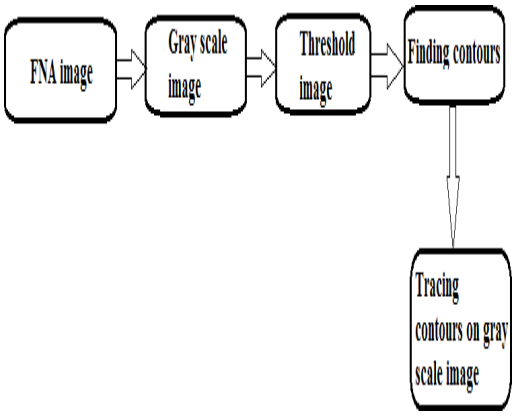
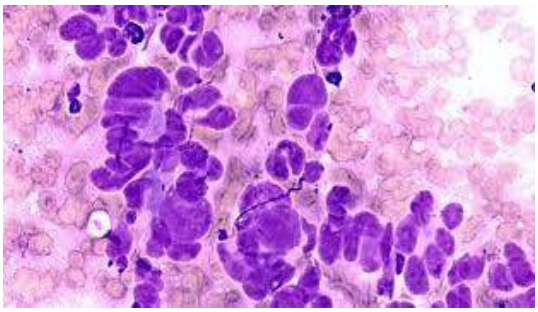


Figure 2: contour tracing



FNA image

Fine needle aspiration is a type of biopsy procedure. In fine needle aspiration, a thin needle is inserted into an area of abnormal- appearing tissue or body fluid. As with other types of biopsies, the sample collected during fine needle aspiration can help make a diagnosis or rule out conditions such as cancer. In here FNA is image that contains cancers cells. On this image, we perform the image processing steps to obtain the contours of the nuclei of the cell.

V. GRAY SCALE IMAGE

Gray-scaling is the process of converting a continuous-tone image to an image that a computer can manipulate. While gray scaling is an improvement over monochrome, it requires larger amounts of memory because each dot is represented by from 4 to 8 bits. GRAYSCALE VS BLACK AND WHITE: In
essence, "gray scale" and "black and white" in terms of photography mean exactly the same thing. However, gray scale is a far more accurate term. A truly black and white image would simply consist of two colors – black and white.

VI. THRESHOLD IMAGE

On the previous process, the FNA image is converted into gray scale image. An image processing method that creates a binary image based on setting a threshold value on the pixel intensity of the original image The thres holding process is sometimes described as separating an image into foreground values Thresholding is a type of image segmentation, where we change the pixels of an image to make the image easier to analyze. In thresolding, we convert an image from color or gray scale into a binary image, i.e., one that is simply black and background values (white).

VII. CONTOUR DETECTION

Having done with the thresholding of the FNA image, we proceed to contour detection. Contours can be explained simply as a curve joining all the continuous points along the boundary, having same color or intensity. The contours area useful tool for shape analysis and object detection and recognition. The nuclei in the thresholded imageareofsameintensity.Socontoursofthesen ucleiareeasilyfoundoutandthey are traced on the gray scale image. The contours have been detected to find various dimensions of the nuclei like radius, area, and perimeter. The above steps were the preprocessing techniques employed for tracing the contours on the gray scale image. Now comes the processes involved for calculation of the dimensions like radius, area, and perimeter that are necessary for calculating the features present in the Wisconsin data set.

VIII. RADIUS CALCULATION

For the calculation of radii of the contours, we approximate circles around them and calculate their radii. It has to be the minimum
enclosing circle to get an accurate result. Then, the mean, standard error, and maximum of the radii values are calculated. For perimeter and area calculation, there is a predefined function for calculation of the length and the area of the contours. Thus using this mean radius, mean perimeter, mean area, standard error of radius, perimeter, and area can be calculated.

IX. Smoothness Calculation

Calculation of „smoothness" feature, involves calculating the local variation of radii of each of the contours, which are not perfect circles. So there will be a variation of radii for each contour. This is calculated easily by a predefined function that calculates local standard deviation. Then, its mean, standard error, and maximum value can be calculated by using known formulas. This gives a measure of how much the nuclei is circular, that is the level of uniformity of the nuclei. Compactness Calculation Compactness is calculated using the following formula perimeter2/area−1.0 As we have already calculated the perimeter and area, this can be calculated easily. This tells us how much dense is the nuclei.

X. Texture Calculation

A few steps following contour detection is required for texture calculation. Texture is basically the standard deviation of gray scale values. Following are the steps to find the grayscale values of thenuclei. After the co ntours aretraced on the grayscale image, we create a masked image of the same shape as the original image and we draw the contours on this masked image The gray scale value of the background is 0 and that of the contour regions. After that the masked image and the gray scale image are bitwise added to have the original nuclei images in this masked image. Then, we get the gray scale values of this image. As the gray scale value for black is zero, whatever value we get is the gray scale value for the nuclei. Then, we find the standard deviation of gray scale values for each of the nuclei. After that the mean, standard error, and maximum value can be calculated using the formula.

XI. Contour Severity

Here again the approximated minimum enclosed circles come into play. The contours have several concave portions. The severity of these concave portions is calculated. The area of this enclosed circle is calculated followed by the area of the contour. Severity is given by the formula Area of the contour / Area of the circle enclosing it. Then it means, standard error and maximum value are calculated. After applying these steps, we are able to extract the required features from the FNA image.

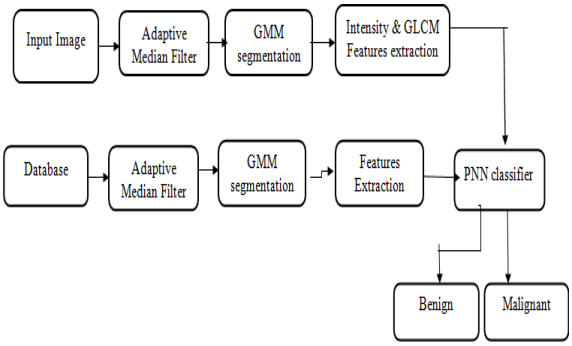


Figure 3: flow chart for machine learning

XII. Process

Thus after obtaining all the features from the image, we put it in the machine learning model. Following is the methodology for building up that model for prediction.

XIII. System description

Nowadays real world databases are highly vulnerable, to noisy, missing and inconsistent data due to their large size. So data preprocessing is necessary before applying any machine learning algorithm. The below steps are involved in machine learning

The breast cancer Wisconsin (diagnostic) data set used here contains 31 columns and 569 entries (rows). It requires data cleaning. The categorical values of the column "diagnosis" have been changed to the binary values, that is "malignant" (cancerous) tumorhasbeenchangedto"1"and"benign" (non-cancerous) tumor has been changed to "0," respectively. The data set is being normalized to bring its mean to 0 and standard deviation to 1. The outliers have been removed to get an evenly distributed data set. Now the data set is ready to apply KNN and SVM algorithm, respectively.
We have tried to apply both these algorithms to the data set and finally find the one with better accuracy. The following subsection contains a brief description of both these algorithms.

XIV. KNN

KNN is a supervised learning technique that deals with classification problems. The algorithm in its initial steps requires choosing a suitable value of "K." The next step of the algorithm states to choose "K" number of nearest neighbors to the new data-point that requires classification. Normally, the nearest neighbors are chosen according to ascending order of the square of the distance of its neighbors to the required data-point. The required data-point is classified to either of the classes depending on which class has more number of neighbors to the data-point. Figure 12 depicts KNN classification for red data-point using K =3 and K =6, respectively.

XV. GMM SEGMENTATION

The segmentation of color image is an important research field of image processing and pattern recognition. A color image could be considered as the result from Gaussian mixture model (**GMM**) to which several Gaussian random variables contribute. In this paper, an efficient method of image segmentation is proposed. Images are represented as arrays of pixels. A pixel is a scalar (or vector) that shows the intensity (or color). A Gaussian mixture model can be used to partition the pixels into similar segments for further analysis. Visualize the distribution of pixel values.

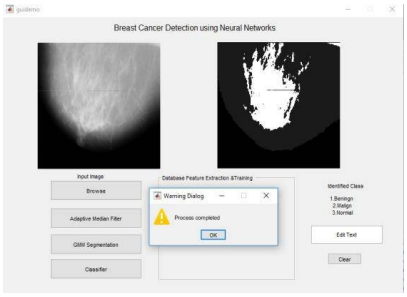


Figure 4: Output of GMM segmentation

XVI. ADAPTIVE MEDIAN FILTER

The Adaptive Median Filter performs spatial processing to determine which pixels in an image have been affected by impulse noise. The Adaptive Median Filter classifies pixels as noise by comparing each pixel in the image to its surrounding neighbor pixels.

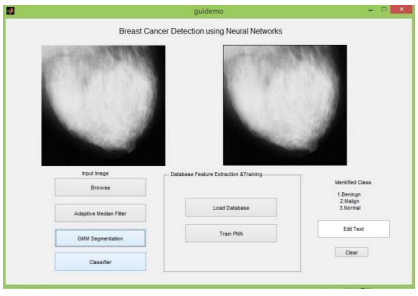


Figure 5: Output of Adaptive median filter

XVII. PNN CLASSIFIER

It is a classifier that maps input patterns in a number of class levels. It can be forced into a more general function approximation. This network is organized into a multilayer feed- forward network with input layer, pattern layer, summation layer, and the output layer. The performance of the proposed structure is evaluated in terms of sensitivity, specificity, accuracy and ROC. The results revealed that PNN was the best classifiers by achieving accuracy rates of 100 and 97.66 % in both training and testing phases, respectively. MLP was ranked as the second classifier and was capable of achieving 97.80 and 96.34 % classification accuracy for training and validation phases, respectively, using scaled conjugate gradient learning algorithm. However, RBF performed better than MLP in the training phase, and it has achieved the lowest accuracy in the validation phase.

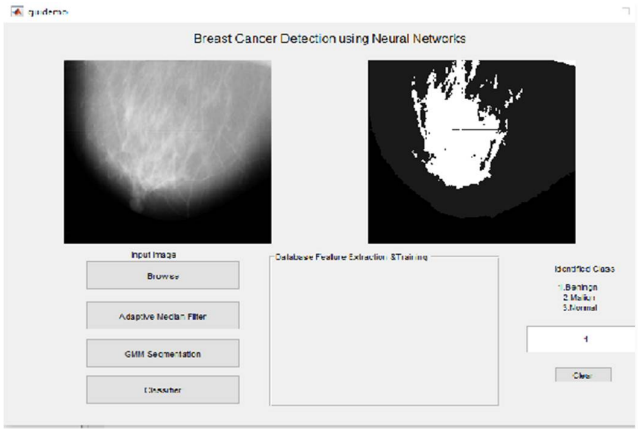


Figure 6: Final output of Breast cancer

XVIII. RESULTS AND ANALYSIS

The model has been trained using 70% of the entries, and the rest of the data set has been used for testing the model. The level of effectiveness of the classification model is calculated by using cross-validation. We present here a comparative study to find out the least number of features that can be used to obtain the best accuracy. In the graph, the results are observed by taking first dominant feature at first, then taking first two dominant features, after those first three dominant features and continue this procedure until the combination of 30 features. Maximum accuracy has been obtained at n=19, and the accuracy is 97.273% using GMM model KNN algorithm has been initially performed on the data set given to find the value of "K" for which there is less error. The graph shows a plot of "K" against "error rate," depicting the least error for the value of "K"=9. Figure represents the plot of accuracy obtained by applying KNN model to the data set obtained each time by varying the total number of features in principal component analysis. The plot represents maximum accuracy of 97.4893% obtained by taking 13 features for data analysis. Here, confusion matrix is used which is a table that is often applied to describe the performance of a "classifier" or classification model on a collection of test data for which the true values are known. The level of effectiveness of the classification model is calculated with the number of incorrect and correct classification in each possible value of the variable being classified in the confusion matrix same accuracy is obtained for both cases of PCA application in case of KNN model. This makes the KNN algorithm more preferable. Moreover, highest accuracy is obtained in KNN using less number of features. Thus, KNN model is finally accepted. From the above comparison table, it is observed that KNN model with PCA value =13 performs to be the best classier.

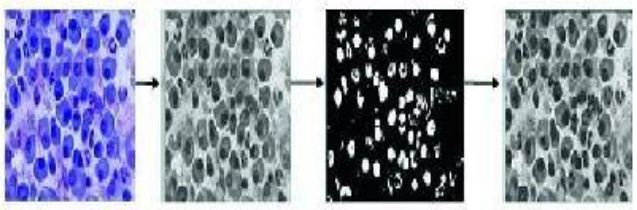


Figure7: FNA image gray scale image threshold image contoured image

XIX. Transition From FNA Image To Contoured Image

The following is obtained for tracing the contours. It is observed that after thresholding the nuclei are segmented. After that finding the contours and tracing them on the gray scale image, the nuclei are distinctly separated. After getting the nuclei, we apply the above steps to obtain the features values. We find a scatter plot where we take y-axis to be set to zero x-axis has been taken to be the feature values obtained. Following scatterplot shows a plot using two values for each feature: one by using the mean of all values of all columns and the second one by using the values obtained by image processing techniques. It is observed that the graph is almost overlapping except for at two points (shown by orange and blue color, respectively). This confirms that the methodology used here is almost accurate. Hence, we finally apply the obtained feature values to the proposed model. This confirms that the affected tumor is "BENIGN," that is "non–cancerous."

XX. Conclusion

Comparing to all other cancers, breast cancer is one of the major causes of death in women. So, the early detection of breast cancer is needed in reducing life losses. This early breast cancer cell detection can be predicted with the help of modern machine learning techniques. The efficiency (97.489%) of this entire model is quite high. Thus in future software can be developed where the FNA slide images can be uploaded and it will automatically process the image to get the feature values and will automatically apply it to the model to predict the result. Thus, it will lead to a very fast diagnosis of breast cancer.

References

[1] Breast Cancer Facts. http://www.nationalbreastcancer.org/breas t-cancer-facts
[2] Fine Needle Aspiration Biopsy. https://www.myvmc.com/investigations/fi ne-needle- aspiration-biopsy-fna/
[3] S.L. Ang, H.C. Ong, H.C. Low, Classification using the general Bayesian network. Pertanika J. Sci. Technol. 24(1), 205–211 (2016)
[4] Wisconsin Breast Cancer Dataset- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29?fbclid=IwAR1seM6TeFo1sHOxjeyH NwElDZDlVyN9mkW2qS Wcl6xi35P2bsIg2A-m8jM
[5] D. Dumitru, Prediction of recurrent events in breast cancer using the Naive Bayesian classifi- cation. Ann. Univ. Craiova-Math. Comput. Sci. Ser. 36(2), 92–96 (2009)
[6] K. Sivakami, Mining big data: breast cancer prediction using DT-SVM hybrid model. Int. J. Sci. Eng. Appl. Sci. (IJSEAS) 1(5), 418–429 (2015)
[7] S.Kharya,S.Agrawal,S.Soni,NaïveBayescl assifiers:probabilisticdetectionmodelforbr east cancer. Int. J. Comput. Appl. 92(10), 0975–8887 (2014)
[8] 8. A. Adam, K. Omar. Computerized Breast Cancer Diagnosis with Genetic Algorithms and Neural Network. fitt.mmu.edu.my/caiic/papers/afzaniCAIE T.pdf
[9] https://dialnet.unirioja.es/descarga/articulo /4036558.pdf 10. D. Dursun, W. Glenn, K. Amit, Predicting breast cancer survivability: a comparison of three datamining methods. Artif. Intell. Med. 34, 113–127 (2005)Breast Cancer Diagnosis Using Image Processing … 127
[10] A.G. Ghuneim. Contour Tracing. http://www.imageprocessingplace.com/do wnloads_V3/ root_downloads/tutorials/contour_tracing _Abeer_George_Ghuneim/intro.html
[11] Mathworks Image Grayscaling- https://www.mathworks.com/help/images/ grayscale-images.html?fbclid=IwAR2Dz1gWFwEts Za-JtXwS1nKq0KCyxtPQQd6- k460zgrT8aB V9M3b9JTWTU
[12] Mathworks Image Thresholding. https://in.mathworks.com/discovery/imag e-thresholding. html
[13] O. Sutton. Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction (2012). http://www.math.le.ac.uk/people/ag153/h omepage/KNN/OliverKNN_ Talk.pdf