

Breast Cancer Detection Using Clustering And SVM

K.S Ghuge

Information technology Sinhgad college of engineering Pune, India ksghuge.scoe@sinhgad.edu

K.S Korabu

Information technology Sinhgad college of engineering Pune, India kskorabu.scoe@sinhgad.edu

Aayushi Somani

Information technology Sinhgad college of engineering Pune, India aayushi.somani@gmail.com

Article Info

Volume 83

Page Number: 2196 - 2205

Publication Issue:

March - April 2020

Abstract

Breast cancer is one of the preeminent impetus of deaths in India. Current Indian data show that one in 22 women develop breast cancer. Out of the two who develop it, one of them would have a higher chance of not surviving it. Breast cancer in India is different from that in the West. Here, it is affecting younger women and more than half of them present themselves in advanced stages. By 2030, breast cancer will cause maximum deaths among women in India than any other cancer. Effective and efficient early detection and classification tools are required to reduce the mortality rates due to breast cancer. Machine learning provides ways to detect and classify these early by finding some recurring patterns in the mammograms. In this research work, fuzzy c-means (FCM) clustering is presented as an approach for performing segmentation and SVM as an approach for performing classification. Image aggrandizing approaches, segmentation, feature adumbration and classification approaches have been delved into and administered on mammographic images obtained from the mini-MIAS repository (Abstract)

Article History

Article Received: 24 July 2019

Revised: 12 September 2019

Accepted: 15 February 2020

Publication: 18 March 2020

Keywords – FCM, SVM, segmentation, mammograms (key words)

I. INTRODUCTION

The aforementioned disease is formed by the calcium deposits i.e. Breast Calcification, developed in the breast tissue. These Calcifications can be benign (i.e. Macro Calcifications) or malignant (i.e. Micro Calcifications). The prognosis of Breast Cancer can be done in three ways: Benign, normal and suspicious calcifications (can be early stage cancer). Breast cancer frequently emerges in lobules, the glands that produce milk or in the ducts, the tube where the milk comes to pass through. The most prevalent class of breast cancer are: Invasive ductal and Invasive lobular carcinoma.

The tumour can be felt as lumps but if is too small then mammograms are usually preferred for

this. The causation of breast cancer is still undiscovered therefore, pronto detection plays a notable role in downsizing the lethality of the disease. The moribund for women due to the aforementioned disease is approximately 1 in every 38, which is about 2.6% of the total estimation of the female population. So, it is paramount to postulate a method which would essentially reduce the mortality rates due to the same. Pronto detection of the tumour will equip the doctors with providing better treatment. Per contra, it is important to provide dependable and concrete conclusions to find a distinguishable solution for different ilk of tumors presenting themselves in patients. In this research work, two classifications are assumed for the type of lumps found i.e. Benign or Malignant. Mammography is a

proficient *modus operandi* to aid in detecting breast cancer.

The Dataset used in this paper is MIAS dataset, which is a very popular and accurate dataset. It has 322 mammographic images. The mammographic images are classified into **Benign** or **Malignant** i.e. **0** or **1** respectively. The characteristics of data present in this dataset are mentioned below in this paper under the Heading VII, topic A.

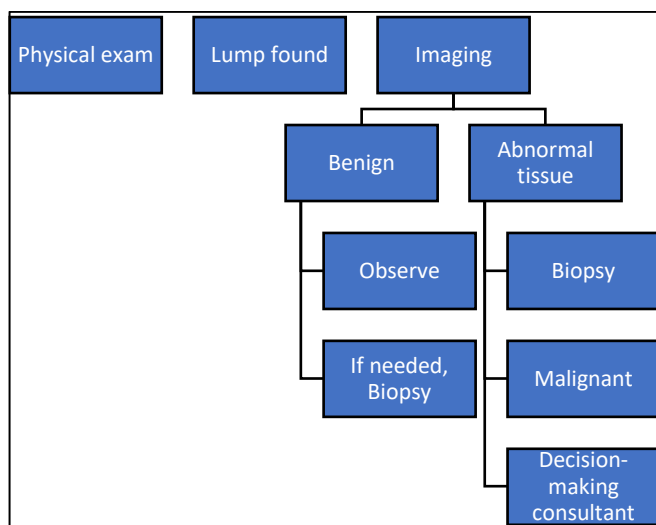


Fig 1: Breast cancer diagnosis

The above figure which is used in this paper for breast cancer diagnosis in this research project. We have used the data available at UCSF's online portal available for breast cancer and complied that in the form of a chart for better understanding. This is how the breast cancer is commonly diagnosed in its relative medical field.

The surpassingly probable symptoms of the aforementioned disease in females are:

- Change in the lineation or size of the mammary glands unequivocally.
- A thickness or lump formation in and around the breast or the underarm region.
- Problems with the nipple region such as: galactorrhea which is discharge from

nipples, tenderness or nipple retraction which is also known as nipple inversion.

- Weathering of the breast tissue more than usual.
- Some evident changes in the usual manner and touch of the breast skin surrounding the areolae and the nippular region like excess warmth, contusion, inflammation.

The primary goal of this project is to discover and extract features from the digital mammograms available in the dataset used. Feature extraction is used to collect and then compose the data from these mammograms. Then, examination with a wad of mammographic images from the database is conducted and these images are segmented through clusters. Subsequently, certain attributes are considered for the features extracted to identify the ilk of the lump formation using a classification algorithm.

In this exhibited work, the detection of the aforementioned disease using clustering as a segmentation technique is studied. According to the flow presented below, after the background noise removal and enhancement of the image, FCM is anointed. Adaptive median filter and histogram equalization methods are exploited to amputate the background noise and perform image aggrandizing, respectively. The most important step of this whole process is the selection of features and then siphoning them out using GLCM (i.e. Gray level co-occurrence matrix). Then, Classification into **Malignant** or **Benign** is concluded using SVM (i.e. Support vector machine).

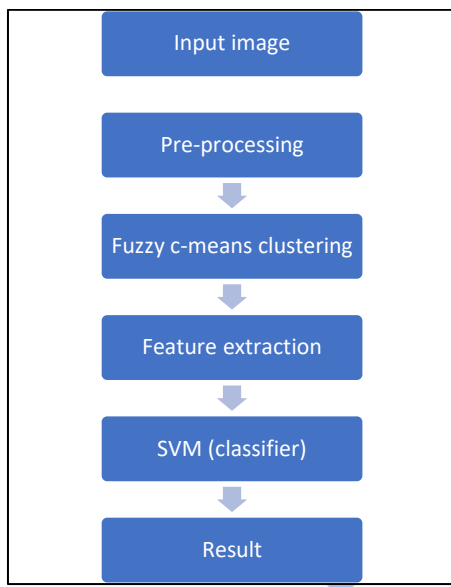


Fig 2: Process of discovering malignancy in the mammary glands

II. LITERATURE SURVEY

Moustapha Mohamed Saleck, Abdelmajid ElMoutaouakkil and Mohammed Mouçouf [1] forwarded a schema for automatized exposé of tumours in digital mammograms. Here, they are using clustering to find the ROI (region of interest) and the fuzzy c-means input data is verified using GLCM to automat this process. The juxtaposition of the results produced by the proffered approach indicated that the proffered approach can proficiently and automatically extract masses from ROI.

Vidya V K and Santo Mathew [2] forwarded a schema for the apprehension of breast cancer by applying probabilistic fuzzy c-means algorithm in Ultrasound images. MATLAB has been used for image processing. It was noted that the error rates were dwindled impressively. This morphological affair conceived effective extraction of the tumour affected regions from or around the breast area. The boundary detection algorithm was applied to distinguish between the benign and malignant masses.

A. Murugan and T. Sridevi [3] forwarded an aggrandized schema for processing of feature

selection, constituting rough sets which are calculated by grouping similar objects together and calculating the next anomaly object from every cluster present using a distance based function formulated. Here, The application and exemplification of the input data is done using some contrivances which are, K-means and FCM (i.e. Fuzzy c-means). The efficiency of FCM(i.e. Fuzzy c-means) surpasses the efficiency of k-means algorithm. This was materialized by employing MATLAB (version 7.12) through the workings of the project.

Prachi Damodhar Shahare, Ram Nivas Giri [4] compared and analysed two separate techniques for breast cancer detection i.e. Artificial neural networks and SVM. Several kernel functions of the Support vector machine (i.e. SVM) operation are employed for the tabulating purposes. The contingent reasoning of both techniques shows Support vector machine (i.e. SVM) employing specific kernels performs better than ANN.

Gehad Ismail, Tarek Gaber, Vaclav Snasel, Mona Soliman, Ahmed Anter, Mona Ali, Aboul Ella Hassanien, Noura Semary [5] forwarded a method to classify the thermograms of the breast region into normal and abnormal. Here, automated segmentation is achieved by employing the NS (i.e. Neutrosophic sets) and optimized F-FCM (i.e. Fast Fuzzy c-mean) algorithm. The database benchmark is employed in this proposed system. The system was appraised using a few quality factors such as: accuracy, precision and recall as well as a comparison with other related work.

Melendez J., Puig D., Valls A., and Torrents-Barrena J. [6] forwarded a method to segregate breast cancer into two experimental types, which are: *Normal* and *Abnormal*, based on mammographic image analysis. An array of Gabor filters is used to determine the standard deviation, mean, kurtosis and skewness characteristics by a window of size four for evaluation and accurate

kernel parameters are used to finally classify using SVM into two classes. Four SVM kernel functions namely, Linear, sigmoid, polynomial and RBF are applied and compared by using a gram matrix.

III. SYSTEM DESIGN

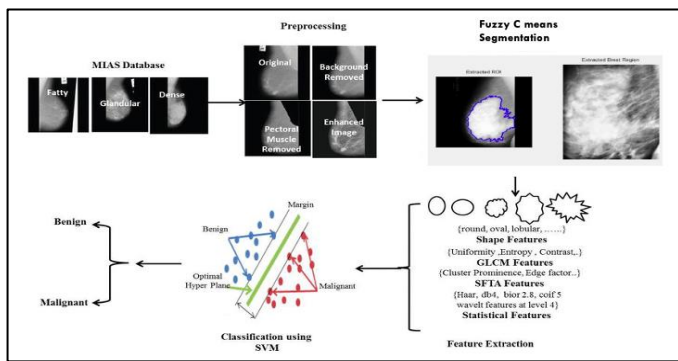


Fig 3: System design

The system design for this research project is shown above in Figure 3. It shows the entire process involved in this system from the data gathering to the classification of the digital mammograms into Benign or Malignant based on the characteristics chosen.

Firstly, the pre-processing is done where the background noise is removed and the image is enhanced by using Adaptive median filter and CLAHE.

Secondly, fuzzy c-means is applied for segmentation.

Thirdly, Features are selected and extracted by using GLCM.

Fourth, Classification is performed using Support vector machine into two classes i.e. Benign and Malignant.

IV. PRE-PROCESSING

Pre-processing is performed to remove the background noise or artefacts. It is the first stage after data gathering, which is treated as an input stream for the pre-processing step. Here, the element factor of the image is augmented by levelling and increasing the contrast to clear the image of any other artefacts. It reduces noise

artefacts, the blurring effect but preserves edges and improves local contrast, Adaptive median filter is used for this. Later on, for certain contrast adjustments being made for the abnormalities or noise in the image to be detectable, we use histogram equalization, where-in, all of the pixel are dislodged by the integral of the histogram for that pixel in the image.

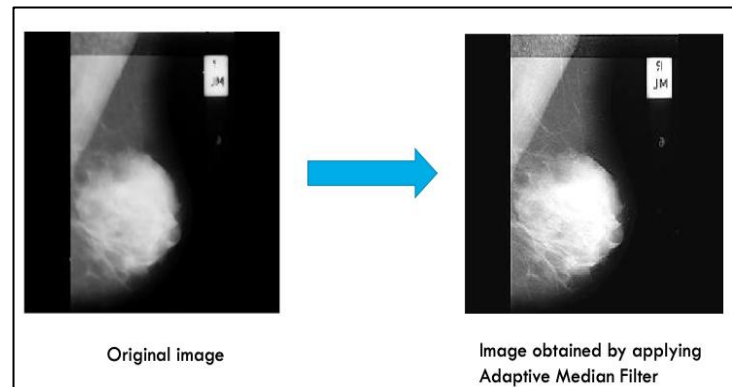


Fig 4: Image filtering using Adaptive median filter

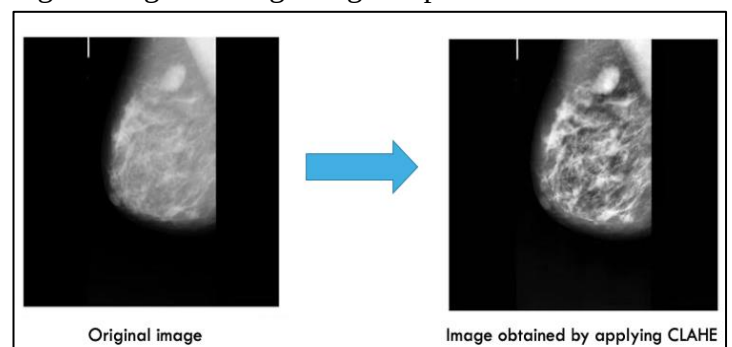


Fig 5: Histogram equalization of filtered image

V. FCM FOR SEGMENTATION

A. Defination

It is the assemblage of a collection of objects grouped together forming clusters where-in, the objects in a particular cluster are more akin to each other than to the objects in other clusters.

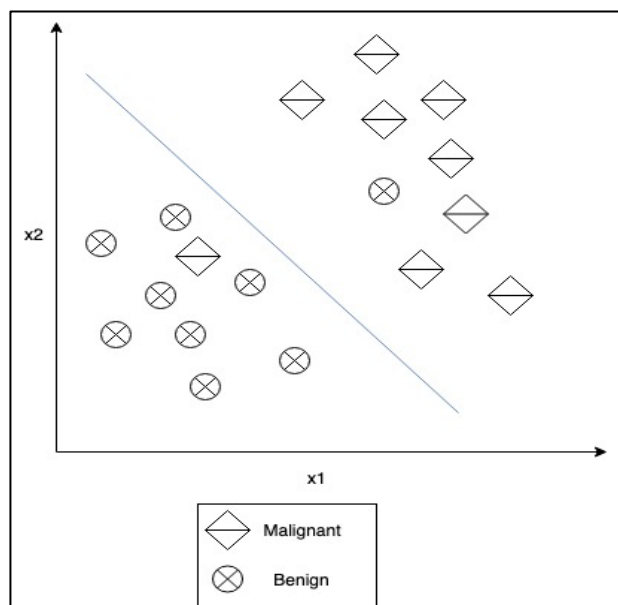


Fig 6: Cluster formation

Here, soft fuzzy c-means clustering is actualized. Soft-FCM unlike hard-FCM is a type of clustering where-in one item can belong to more than one clusters but an object will be more analogous to one cluster than the rest. The minimization of the function below explains this:

$$X_m = \sum_{i=1}^n \sum_{j=1}^k D_{ij} ||Y_i - C_j||$$

Here,

- C = Centroid of the j^{th} cluster and
- $Y = i^{\text{th}}$ data point.

B. Distance Matrix

Clustering gives good accuracy and is easier to implement too. In this, the data points are divided using measures which forms clusters i.e. groups which have mostly alike data points. Then, a centroid is calculated for each cluster and no. of clusters are defined. When the distance span between the data points and the centroid doesn't change much then it stops.

It is represented by a $[m \times n]$ matrix D . where m is the number of clusters at the result end and n is

the number of data points from the original data set. Each entry field in the D matrix is denoted by D_{ba} where,

$$b \in \{1, \dots, m\},$$

$$a \in \{1, \dots, n\}$$

$$D = \begin{bmatrix} D_{11} & D_{12} \dots & D_{1n} \\ \vdots & \ddots & \vdots \\ D_{m1} & D_{m2} \dots & D_{mn} \end{bmatrix}$$

If, $D_{ba} = 1$, the data point a is in cluster. else, $D_{ba} > 0$, the data point isn't in cluster

C. Algorithm

1. Do a first random partition of clusters and appoint values to D_{ba} .
2. Repeat.
3. Enumerate the centroids or every cluster by adopting the aforementioned minimization function for FCM.
4. Continue to update the random partition made earlier by distance matrix mentioned above.
5. Pursue this until the centroids don't alter much (eg: error value is below a aforementioned threshold or change in D_{ba} below a specified threshold).

VI. FEATURE EXTRACTION

In this step, features obtained in the segmentation phase are extracted using several statistical features, which are:

- Energy: The quantification of the regularity obtained in an image.
- Entropy: Entropy gives the measure of randomness.
- The value is paramount when facets from the co-occurrence matrix are perpetual.

- **Homogeneity:** Homogeneity measures the contiguity of apportioning of the elements through GLCM approach. Its limits are from 0 to 1.
- **Correlation:-** Correlation measures the pertinence between the reference and the neighbouring pixel.
- **Contrast:-** Contrast gauges the intensity fluctuation amid reference pixel and neighbouring pixel.

We use these statistical components in the feature extraction process in the mammographic images by using GLCM (i.e. Gray level Co-occurrence matrix). The a foreknown **GLCM** functions outlines the consistency of an image by assessing the frequent nature of the pixel pairs with certain minutiae and in a aforementioned spatial accord which occurs to an image, to make a **GLCM matrix**, henceforth the statistical characters are culled in distinction to this matrix. In order to characterize the textual features of an image, the direction of analysis is selected, which can be as follows:

- Vertical : 90 degrees
- Horizontal : 0 degree
- Diagonal:
 - ❑ Bottom left to top right(i.e. -45 degrees)
 - ❑ Top left to bottom right(i.e. -135 degrees)

Denoted as $A_0, A_{45}, A_{90}, A_{135}$ respectively.

Example: $A(i,j)$

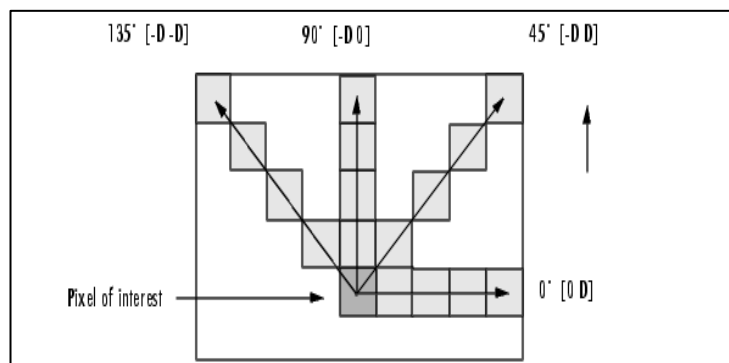


Fig 7: GLCM directional analysis

VII. CLASSIFICATION

SVM is a Panglossian pattern for classification approach which was proposed not long ago under the aegis of Vapnik and Cortes in 1995. The desideratum of this algorithm is that it works at attenuating the upper hem of the “*generalisation error*” by the virtue of augmentation the margin middling by separating the hyperplane and it’s data points[6]. This algorithm minimizes the marginalized error by augmenting the rift between the hyper-plane and the data points. SVM has been an exceedingly propitious algorithms to classify breast cancer in mammograms.

A. Classification

In this research work, SVM (i.e. Support vector machine) is mobilized to segregate the data points into two ilks namely, *Benign* and *Malignant*. The intuition used for SVM is *special linearly seperable* classification case. SVM finds the largest margins between the closest data points for each class and linearly seperates them. This is commonly known as “Maximum margin hyperplane (MMH)”.

The classification is performed on the basis of the following characteristics of the breast tissues:

1. Tone of the rearing tissues:
 - a. D – Dense and glandular tissues
 - b. F - Fatty tissues
 - c. G – Fatty and glandular tissues

2. Category of aberrancy present:
 - a. Calcification
 - b. Masses, types:
 - i. Speculated
 - ii. Delineate/circumscribed
 - iii. Other, murky masses
 - c. Asymmetry
 - d. Architectural distortion
 - e. Normal
3. Austerity of the lump present:
 - a. M - Malignant

B – Benign

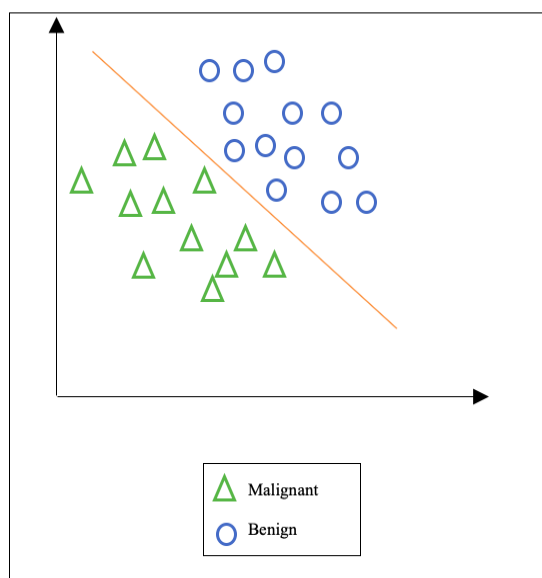


Fig 8: SVM linear classifier

B. Algorithm

1. After features are extracted, we get the labelled data.
2. +1 is used for benign and -1 for malignant images. After the classes are defined and the labels are assigned to the training data, SVM reads input data one by one and plot them in the two-dimensional feature environment.
3. After plotting the data points, we have to find an optimal hyperplane which will classify the data correctly.

4. Line with maximum margin from the data points is selected. Hyper-plane with maximum margin will classify the data into benign and malignant.

VIII. RESULT AND ANALYSIS

```
filter_image = cv2.medianBlur(image2, figure_size)
plt.figure(figsize=(11,6))
plt.subplot(121), plt.imshow(image2, cmap='gray'),plt.title('Original')
plt.xticks([], plt.yticks([]))
plt.subplot(122), plt.imshow(filter_image, cmap='gray'),plt.title('Median Filter')
plt.xticks([], plt.yticks([]))
plt.show()
```

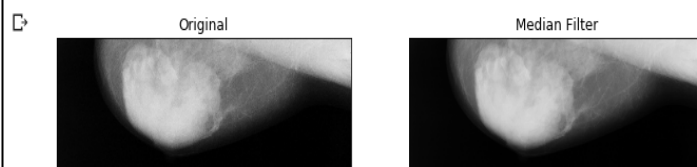


Fig 9: Median filter applied

```
[18] #CLAHE conversion
clahe = cv2.createCLAHE(clipLimit=2.0, tileGridSize=(4,4))
new_image = clahe.apply(filter_image)
plt.figure(figsize=(11,6))
plt.subplot(121), plt.imshow(cv2.cvtColor(image, cv2.COLOR_HSV2RGB)),plt.title('Median filter')
plt.xticks([], plt.yticks([]))
plt.subplot(122), plt.imshow(new_image),plt.title('CLAHE')
plt.xticks([], plt.yticks([]))
plt.show()
```

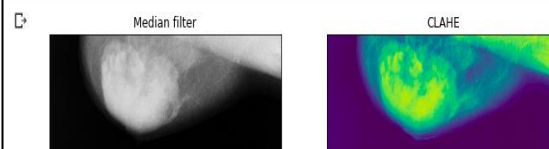


Fig 10: CLAHE applied

```
[19] # CLAHE gray scale conversion
clahe = cv2.createCLAHE(clipLimit=2.0, tileGridSize=(4,4))
grey_image = clahe.apply(new_image)
plt.figure(figsize=(11,6))
plt.subplot(121), plt.imshow(new_image, cmap='gray'),plt.title('Clahe')
plt.xticks([], plt.yticks([]))
plt.subplot(122), plt.imshow(grey_image, cmap='gray'),plt.title('Clahe grey scale')
plt.xticks([], plt.yticks([]))
plt.show()
```

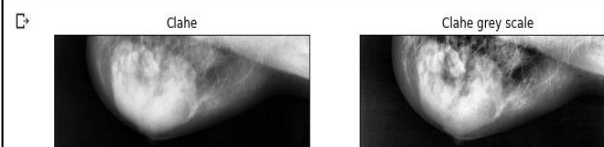


Fig 11: CLAHE grey scale conversion

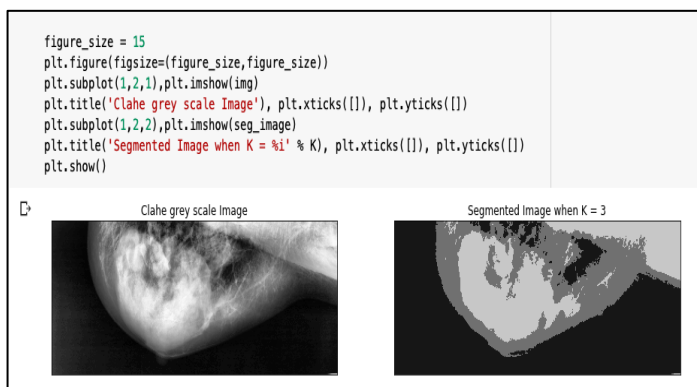


Fig 12: Image segmented using fuzzy c-means

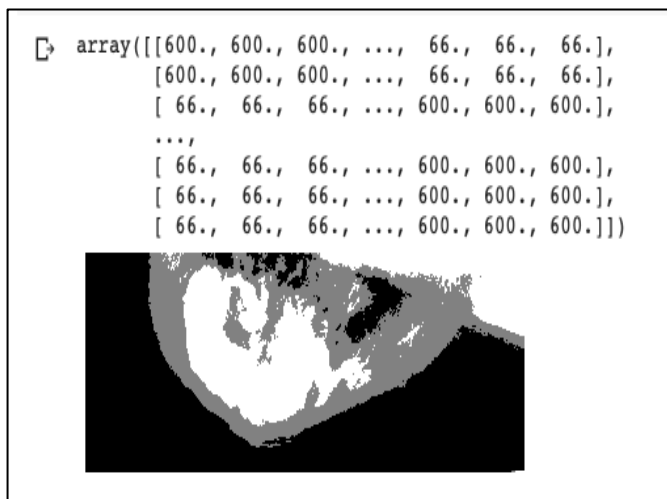


Fig 13: Features extracted from segmented image

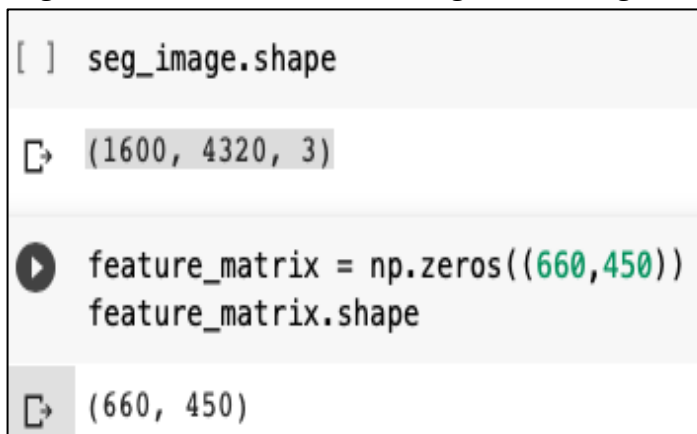
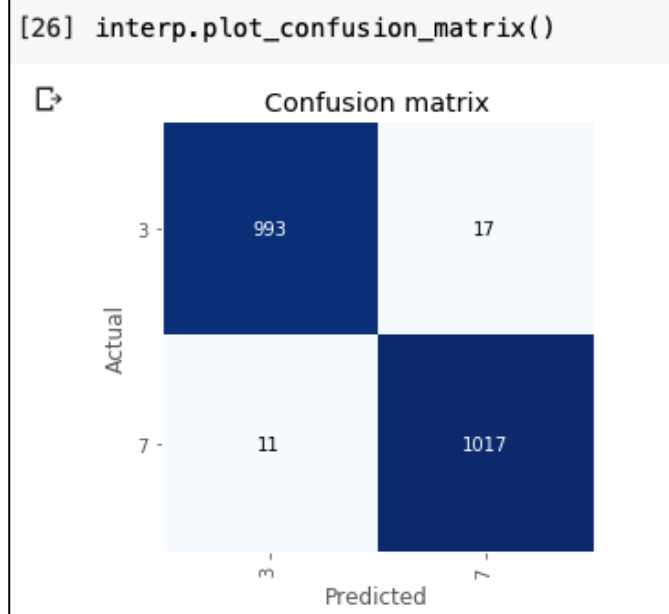


Fig 14: Feature matrix

Fig 15: Confusion matrix

The indicators for measurement of the system:

$$1. \text{ Accuracy} = \frac{TN+TP}{TOTAL}$$



$$2. \text{ Precision} = \frac{TP}{TP+FP}$$

$$3. \text{ Recall} = \frac{TP}{TP+FN}$$

$$4. \text{ Specificity} = \frac{TN}{TN+FP}$$

INDICATOR	RESULT
Accuracy	0.986
Precision	0.9835
Recall	0.9892
Specificity	0.9831

Table 1: Indicators found using the confusion matrix

The above table is the indicator results calculated using the confusion matrix in Fig 9, made using the dataset mentioned above i.e. MIAS. It was calculated using pytorch and its library CUDA and the images were displayed using the library Image Data Bunch.

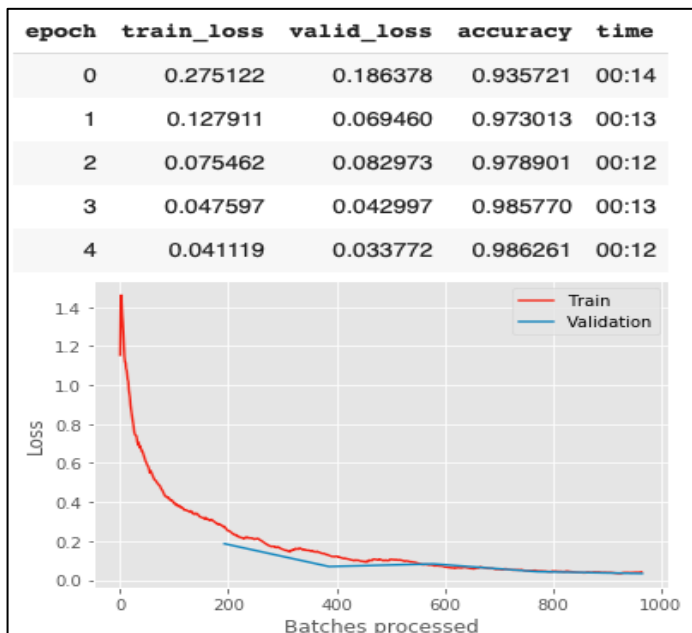


Fig 16: Training and validation results based on accuracy and time

These results are based on 5 epochs starting from 0 to 4, which depict the training loss, the validation loss, accuracy and the time taken for that epoch.

The dataset was divided into training, testing and validation dataset. The validation dataset was used to find the accuracy of the training dataset. The training data is tested against the validation data to find its accuracy, time and the loss obtained from each dataset.

It was found that the best filter to use for approaching the best pre-processed image in this case was Median filter, the image appeared with much less abrasions than with Mean or Gaussian filter. CLAHE has been established to be quite accurate and performative in terms of conversion to the respective histogram equalized image. The contrast, entropy and other features were taken in account for the same. Later on, features were extracted from the then segmented image and a matrix of these features was created for better classification, which was performed using the support vector classifier. The accuracy achieved using the same classifier has been mentioned above along with several other measurement indicators.

IX. CONCLUSION

In this research work, it is established that SVM and clustering together would get a more accurate result and even the time taken to establish that would be less. This would significantly help in minimizing the mortality rates by a significant number.

X. REFERENCES

- [1] Moustapha Mohamed Saleck, Abdelmajid ElMoutaouakkil and Mohammed Mouçouf, "Tumor detection in mammography images using fuzzy c-means and GLCM texture features", 2017 14th International Conference on Computer Graphics, Imaging and Visualization.
- [2] Vidya V K and Santo Mathew, "An accurate method of breast cancer detection from ultra sound images using probabilistic fuzzy clustering algorithm", International Conference on Communication Systems and Networks (ComNet) | 21-23 July 2016.
- [3] A.Murugan and T.Sridevi, "An enhanced feature selection method comprising rough set and clustering techniques", IEEE 2014.
- [4] Prachi Damodhar Shahare, Ram Nivas Giri, "Comparative analysis of artificial neural network and support vector machine classification for breast cancer detection", IRJET Volume: 02 Issue: 09, Dec-2015.
- [5] Tarek Gaber, Gehad Ismail, Ahmed Anter, Mona Soliman, Mona Ali, Noura Semary, Aboul Ella Hassanien, Vaclav Snasel, "Thermogram breast cancer prediction approach based on neutrosophic sets and fuzzy c-means algorithm", 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).
- [6] Torrents-Barrena, J., Puig, D., Melendez, J., & Valls, A. , "Computer-aided diagnosis of breast cancer via Gabor wavelet bank and binary-class SVM in mammographic images", Computer-aided diagnosis of breast cancer via Gabor wavelet bank and binary-class SVM in mammographic images. Journal of Experimental & Theoretical Artificial-Intelligence.
- [7] Mihir Sewa, Priyanka Vaidya, Chien-Chung Chan, Zhong-Hui Duan, & Valls, A., "SVM Approach to Breast Cancer Classification", Second International Multisymposium on Computer and Computational Sciences.

- [8] Maleika Heenaye and Mamode Khan, “Automated Breast Cancer Diagnosis Using Artificial Neural Network (ANN)”, 3rd Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS-2017).
- [9] Vikas Kumar Saubhagya, Asha Rani and Vijander Singh, “ANN based Detection of Breast Cancer in Mammograph Images”, 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES-2016).
- [10] Chaitanya Varma and Omkar Sawant, “An Alternative Approach to Detect Breast Cancer using Digital Image Processing Techniques”, International Conference on Communication and Signal Processing, April 3-5, 2018, India.
- [11] Punam S. Pawar and Dharmaraj R. Patil, “Breast Cancer Detection Using Neural Network Models”, 2013 International Conference on Communication Systems and Network Technologies.
- [12] Karmilasari, Suryarini Widodo, Matrisnya Hermita, Nur Putri Agustiyani, Yuhilza Hanum, “Sample K-Means Clustering Method for Determining the Stage of Breast Cancer Malignancy Based on Cancer Size on Mammogram Image Basis ”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 3, 2014.