International Journal of
Science and Engineering Investigations vol. 9, issue 106, November 2020
Received on November 21, 2020 ISSN: 2251-8843

Computer Aided Diagnosis of Breast Cancer in Human Mammogram Using Support Vector Machine

A. M. Ikotun[1], D. K. Opiarighodare[2], A. P. Adelokun[3], F. M. Okikiola[4], O. N. Lawal[5]
[1,2,3,4,5]Department of Computer Technology, School of Technology, Yaba College of Technology, Yaba, Lagos State, Nigeria
([2]opiadon@yahoo.co.uk)

Abstract-This study presents a Computer Aided Detection (CAD) system which automatically detects breast cancer in human mammographic images. The system aimed to minimize the problems of false detection of breast cancer by a technique that detects and classifies cancer in human mammograms with Support Vector Machine (SVM) in combination with Gray Level Co-occurrence Matrix (GLCM). This detection and classification technique involves four basic stages besides data acquisition. These include image preprocessing, image segmentation, features extraction and classification. The system applies median filter for noise removal, and Contrast Limited Adaptive Histogram Equalization (CLAHE) for enhancement in preprocessing, watershed transform in segmentation, Gray Level Co-occurrence Matrix (GLCM) in feature extraction, and Support Vector Machine (SVM) in two-phase classification. The performance of the two-phase SVM classification system was evaluated using sensitivity, specificity and total accuracy which yielded 96.67%, 98.33% and 97.5% respectively. This shows a more pronounce True Positive (TP) result, indicating that the two-phase SVM classification has great potentials in the detection of breast cancer with human mammogram than most of the other existing techniques.

Keywords- *Computer Aided Detection (CAD), Mammogram, Contrast Limited Adaptive Histogram Equalization (CLAHE), Gray Level Co-occurrence Matrix (GLCM), Support Vector Machines (SVM)*

I. INTRODUCTION

Breast cancer is considered to be one of the leading and most dangerous diseases that can cause severe illness and fatality resulting death among female folks in many countries of the world (Mohammad T., 2017). Mammographic images are female breast region X-Ray images displaying points with high intensities density which are likely potential tumours (Naser S. and Mohammad R. H., 2019). In mammogram investigation for breast cancer, the tumours and masses are seen as the form of dense regions in the film. Cancer tumours are classified into Benign and Malignant depending on the severity. Tumours that are slow growing and less harmful are classified as Benign while those that grow fast and affect surrounding tissues are classified as Malignant (Ireaneus A. Y. and Rejani S. T. S., 2009). Typically, benign mass is characterized with a round, smooth and well circumscribed boundary while a malignant tumor is usually characterized by a speculated, rough, and blurry boundary. The malignant cells are originally created from milk glands of the breast and they are classified into different groups according to their unusual progress and capability to affect other normal cells. The capability of affecting means whether these malignant cells affect only the local cells or can spread throughout the entire body. However, the effect of spreading these malignant cells throughout the whole body of the patient is known as metastasis (Meenalosini S. et al., 2012).

According to Mohammad et al., (2018) cancer masses usually cause uncontrolled multiplication in any area of the human body and many people have seen their loved ones die due to the cancer disease. A cancerous tumor grows and multiplies out of control, growing as large as 2mm or more every three months and spreading to other parts of the body and destroying the surrounding healthy tissue.

Since there is no clear cause of breast cancer, early detection is of utmost importance in its treatment and/or management, which can be done through various techniques. The use of electromagnetic waves on human body for the detection of cancer has been an ongoing research area over the years. Some years back, microwave system was the possible solution and X-rays were also used to detect breast cancer, however these have negative side effects on patients' breast tissues. (Chithra and Dhivya, 2017).

Many diagnostic methods such as mammography, magnetic resonance imaging (MRI), Ultrasonography, positron emission tomography (PET) and Biopsy Investigators have been researched upon in breast cancer early-stage diagnoses. Breast cancer early detection by the use of mammogram is one of the important methods, however, it has some major drawbacks. Firstly, it is not effective for subjects that are under the age of 40 years. Secondly, in dense breasts, it is less sensitive to small tumours that are less than 1mm, about 100,000 cells) and thirdly, it does not provide any indication of eventual disease outcome. The most attractive alternative to mammogram is the magnetic resonance imaging (MRI). The MRI test is carried out to confirm the existence of tumour. However, a skin infection could develop at the place of injection, or sometimes, a patient could develop an allergic reaction to the MRI contrasting agent.

Besides the screening techniques above, breast biopsies are usually carried out in order to differentiate benign from cancerous tissues. However, this procedure is expensive and requires the service of trained personnel and time just like the constrast-enchanced digital mammography which involves high radiation levels. More recent techniques which were devised to detect the architectural distortion and mass in mammogram images in order to improve on the aforementioned methods include Gabor Wavelet in combination with Adaptive Neuro–Fuzzy based classification (Ragupathy U. S. and Saranya T., 2012), Artificial Neural Network (ANN) (Deepa S. N. and Aruna Devi B., 2011), Gabor Wavelet in combination with Discrete Wavelet Transform (DWT) (Salve S. M. et al., 2013), Gabor Wavelet in combination with Support Vector Machine (SVM) (Snehal A. M. and Kulhalli K. V., 2015), Law's Texture Energy Measure and Neural Networks (Setiawan A. S. et al., 2015), New Asymmetric Fractal Features (Beheshti S. M. A. et al., 2016), Local binary Patterns and Radial Lengths through an Exhaustive Evaluation Framework (Chatzistergos S. E. et al., 2018) as well as combination of several methods jointly investigated by Naser Safdarian and Mohammad Reza Hediyehzadeh in 2019 etc.

Setiawan and other researchers carried out a study on features of mammographic images which were obtained from MIAS database in 2015. They applied Law's Texture Energy measure and Neural Network, and the total accuracy attained is 93.90%. Beheshti and others applied the new asymmetric fractal technique in 2016 in order to detect the abnormalities in the mammogram images. 168 images were carefully selected from MIAS database for the technique by a radiologist, alongside masses ascertained by biopsy and the total accuracy obtained is 94.01%. Chatzistergos and others put forward a technique for classification of mammographic images obtained from MIAS and Digital Database for Screening Mammography (DDSM) in 2018. The technique involves a combination of local binary pattern operators and radial lengths, and the total accuracy attained is 82.54%. And one of the most recent study conducted by Naser Safdarian and Mohammad Reza Hediyehzadeh in 2019 applied a combination of several methods with Support Vector Machine (SVM), using mammographic images from DDSM database, and the total classification accuracy attained is 97% (plus or minus 4.36%).

However, this study is aimed at solving the problems of false detection of breast cancer by a technique that can be used for early detection and classification of cancer in human mammograms using support vector machine (SVM) in combination with Gray Level Co-occurrence Matrix GLCM). The study is important because accurate detection of breast cancer disease in the early stage is extremely essential for fast recovery, and to avoid the death probability. This will in turn, increase the chances of a successful treatment and ensure accurate interpretation of mammograms for detection of suspicious lesions and classification. It will also help in the eradication of unwanted biopsy and reduce stress for women with the disease as well as the time required in reading mammography.

II. MATERIALS AND METHOD

The CAD system developed uses mammogram images as input. The digital mammograms used here were acquired from the Mammogram Image Analysis Society (MIAS) database. The mammographic images fetched from the database are with the "truth" markings on the locations of any abnormalities that may be present. Besides the data acquisition, this system of detection and classification is divided into four major stages including preprocessing, segmentation, feature extraction and classification.

A. Preprocessing

Mammogram images have different artifacts as well as noises in their background. There are pectoral muscles also in the object area, and all these constitute the unwanted portions of the images for the texture analysis. Such portions make the entire mammographic images unsuitable for feature extraction which in turn, causes inaccuracy in classification. To remove these unwanted portions, the mammogram images were subjected to preprocessing, cropping operations to extract the regions of interests (ROIs) where the abnormalities are present.

Generally, it is difficult to interpret mammogram images and this make preprocessing a necessity in order to improve images' quality and features that will be extracted. The preprocessing stage consists of two main phases. In this study, the first phase was carried out by removing the background information and impulse noise from the mammogram images using an adaptive median filters with high denoising ability and efficient computational time. The main steps followed in the first phase of preprocessing are summarized in the steps below:

Step 1: Reading and displaying of image.

Step 2: Adding of noise to the image

Step 3: Filtering the noisy image using an average filter and displaying the results.

Step 4: Using of adaptive median filter for filtering the noisy image and displaying the results.

The second phase of the preprocessing involves enhancing the contrast of interest areas, and this was done by the use of Contrast Limited Adaptive Histogram Equalization (CLAHE) technique. Enhancing the mammogram images helps to brighten the region of interest (ROI) and makes it easier to identify key features for extraction. The main steps followed in this second phase of the preprocessing are summarized below:

Step 1: Load Image using the 'imread' function.

Step 2: Read in two grayscale images: pout. and tire. Also read in an indexed RGB image: shadow.tif.

Step 3: Resize the image to 256 x 256 pixels using 'imresize' function.

Step 4: Enhance the gray scale image using adaptive histogram.

Step 5: Display the result.

B. Image Segmentation

Segmentation is the process of partitioning a picture into a semantically interpretable region. Image segmentation is typically used to locate objects and boundaries in images. The result of segmentation is a set of regions that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region is similar with respect to some characteristics or computed properties, such as colour, intensity or texture. In this study, the marker-controlled watershed algorithm was used for segmentation of mass. The algorithm started by reading the image and using the gradient magnitude as the segmentation function. The Watershed Transform produced a binary image with watersheds regions which were assigned 1(black) while regions surrounded by dams were assigned 0 (white). The algorithm below was used for the image segmentation;

Step 1: Read in the Grayscale image.

Step 2: Use the gradient magnitude as the segmentation function.

Step 3: Mark the Foreground Objects.

Step 4: Compute Background Markers.

Step 5: Compute the Watershed Transform of the Segmentation Function.

Step 6: Visualize the Result.

C. Feature Extraction

The extraction and selection of features from a mammogram image is of great significance in the classification of breast cancer with mammographic images. This is because conventional mammographic images are highly textured and complex, which in turn makes interpretation of the images difficult even when the images consist of spatial resolution of x-ray that is in the order of few microns which allows visualization of the masses. In this study, Gray Level Co-occurrence Matrix (GLCM) approach was applied for the extraction of discriminating features used in classification. The GLCM represents the various image classes. The texture and statistical features selected are mean, standard deviation, contrast, correlation, energy, homogeneity and entropy.

During the classification, the image properties obtained from the extracted features were used in comparing unknown sample image features for correct classification. The following describe some of the features and equations that were used for the extraction:

- Mean: It is the average value of intensity of the Image. It is defined as:

$$\mu = \sum_{I=0}^{L-1} Zi\ P(Zi) \tag{1}$$

- Standard deviation: It is the square root of the variance. The standard deviation is the estimate of the mean μ square deviation of gray pixel value P (Zi) from its mean value. It is defined as:

$$SD = \sqrt{\sigma^2} = \sqrt{\sum_{I=0}^{L-1} (z_i - \mu^2) p(z_i)} \tag{2}$$

- Correlation: The operation called correlation is closely related to convolution. In correlation, the value of an output pixel is also computed as a weighted sum of neighbouring pixels. It is defined as:

$$C = \frac{\sum_i \sum_j p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3}$$

- Energy: This gives the measure of the uniformity. It can be calculated as sum of the element's square of pixel values. It is defined as:

$$E = \sum_{I=0}^{L-1} [P(Zi)]^2 \tag{4}$$

Homogeneity: The closeness of the distribution of elements in the GLCM to the GLCM diagonal is measured by homogeneity. It is calculated as:

$$\text{Homo: sum(sum(p(x,y)/(1+[x-y])))} \tag{5}$$

- Entropy: Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. It is defined as:

$$E = -\sum_{i=0}^{L-1} P(z_i) \log_2 [p(z_i)] \tag{6}$$

D. Classification with the use of Support Vector Machine

The expressed features which were extracted from the detected masses were transformed into vectors in storage. By these vectors, feature matrix was created which was applied as a feature set for the input of the classification system. Support Vector Machine (SVM) was used as the classifier. The classification is in two stages and this made the SVM to be trained twice for proper classification using the three attributes of interest (normal, benign and malignant). The first classification stage classified mammogram images into either normal or abnormal while the second classification stage classified the abnormal mammogram images into either benign or malignant. The system flowchart including the two-stage classification approach is illustrated in figure 1.

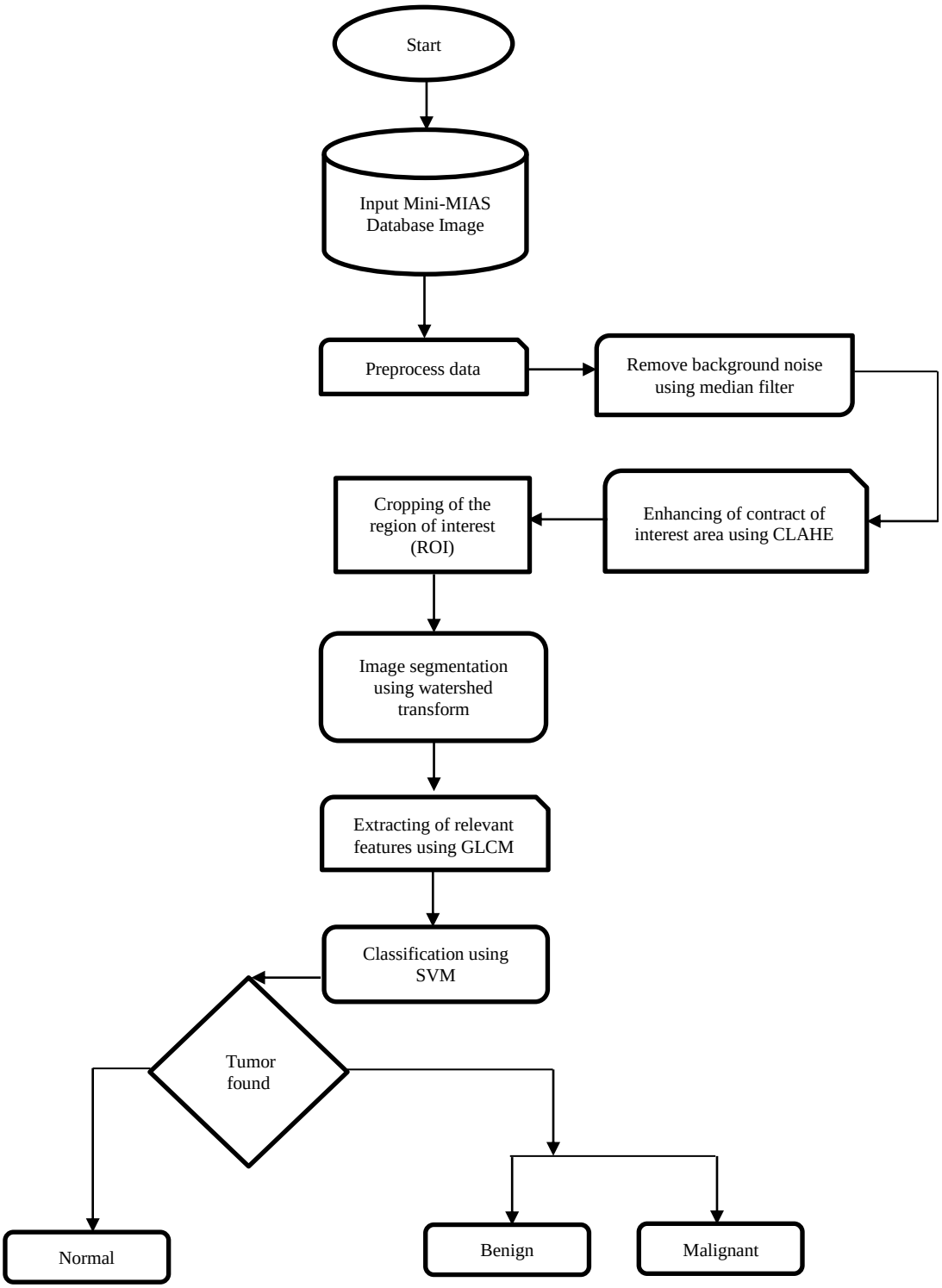

Figure 1. System Flowchart

III. RESULTS AND DISCUSSION

The mini mammographic images database provided by Mammographic Image Analysis Society (MIAS) served as the data source for all the mammogram images used in this study. 322 images (Medio-Lateral Oblique (MLO)) which represents 161 bilateral pairs at 50-micron resolution in "Portable Gray Map" (PGM) format with the associated truth data make up the entire collection. The samples of results of the first stage of preprocessing and the second stage which is segmentation are illustrated as shown below:

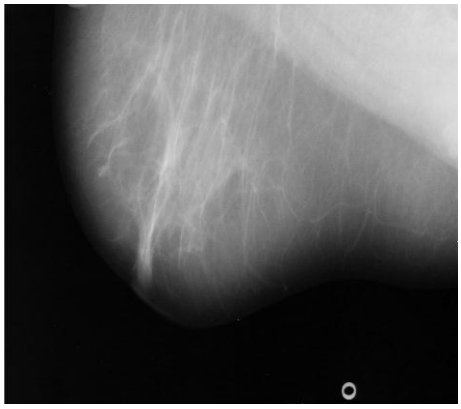

Figure 2. Original Mammogram Image Sample Derived before Processing

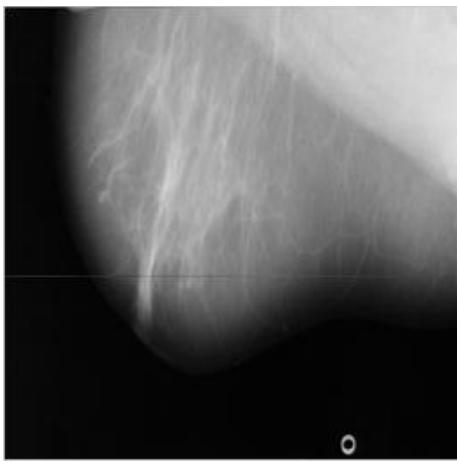

Figure 3. Resized Image (256 x 256)

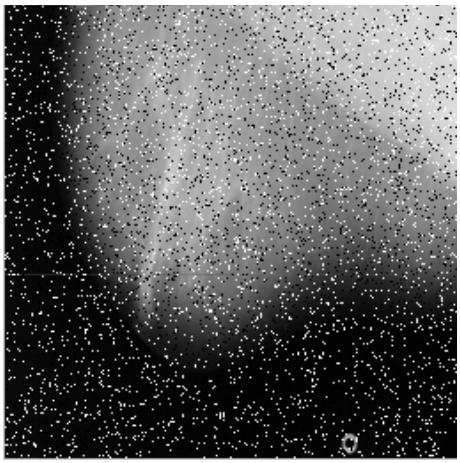

Figure 4. Mammogram Image with Noise

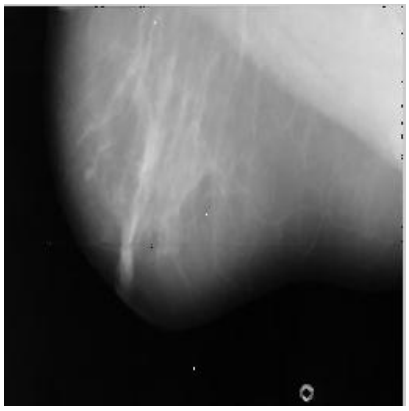

Figure 5. Mammogram Image Output by Median Filter

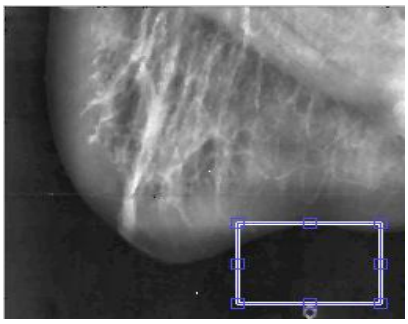

Figure 6. Enhanced Mammogram Image Output by CLAHE

Furthermore, the image enhancement which was carried out by the use of CLAHE technique was followed by a cropping operation. This operation was done on the adapthisteq mammogram images to extract the regions of interests (ROIs) and the result is illustrated in figure 7:

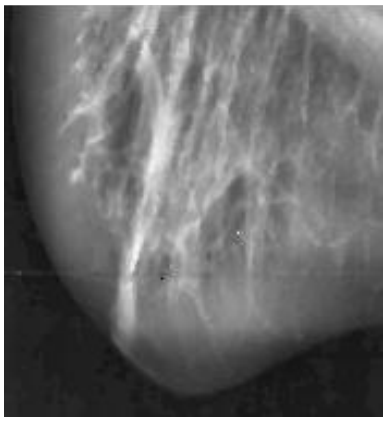

Figure 7. Mammogram Image Region of Interest (ROI)

Figure 8. Binary Mask (Segmentation 1)

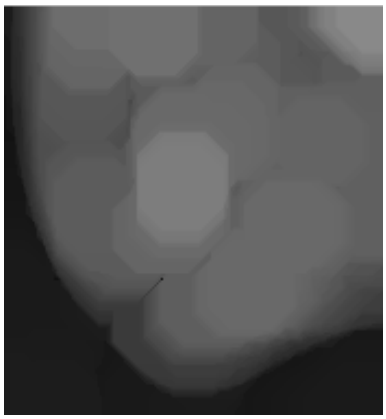

Figure 9. Gradient Magnitude (Segmentation 2: Watershed Transformation)

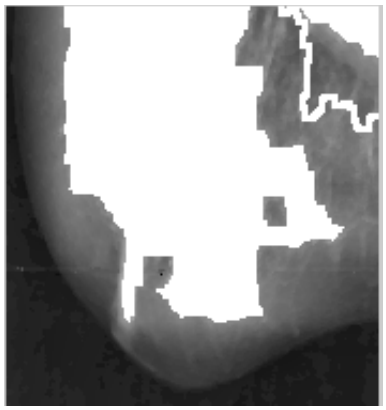

Figure 10. Image Dilation (Segmentation 3: Watershed Transformation)

For extraction of features, the relevant features extracted with the aid of Gray-level co-occurrence matrix (GLCM) technique are mean, standard deviation, contrast, correlation, energy, homogeneity and entropy etc.. A segment of the selected features extracted during the feature extraction stage is shown in table 1.

TABLE I. SEGMENT OF SOME SELECTED FEATURES FROM EXTRACTION

Mean	Standard Deviation	Contrast	Correlation	Energy	Homogenity	Entropy
95.0112	103.7595	0.1524	0.9751	0.1499	0.9268	5.057
22.5353	0	0.0038	0.9273	0.9542	0.9981	0.6012
83.7942	95.4584	0.6437	0.9503	0.2025	0.8722	3.249
96.9715	102.4539	0.2322	0.9558	0.1338	0.8911	5.1739
104.2163	97.7814	0.183	0.957	0.1655	0.9119	5.4992
95.9327	103.1327	0.1802	0.9685	0.1475	0.9173	5.3393
104.6152	98.8846	0.2117	0.9528	0.1428	0.8996	5.1922
151.1452	58.7326	0.2239	0.8655	0.2212	0.9073	3.9115
52.618	58.4107	0.4562	0.7316	0.4049	0.8821	2.9694
104.6687	96.6057	0.245	0.9471	0.1312	0.8886	5.5878
111.7753	89.4789	0.2222	0.9532	0.1439	0.8946	5.3892
52.4705	58.5718	0.0784	0.9678	0.2671	0.9609	5.0398
97.4312	100.9086	0.1768	0.9679	0.1391	0.9162	5.6796
101.5071	100.0004	0.2161	0.9598	0.1246	0.8992	5.7466
97.5143	98.7993	0.2172	0.9595	0.1339	0.8988	5.0896
104.9496	101.4972	0.2187	0.9654	0.1231	0.8971	5.3785
102.8366	99.883	0.1994	0.9696	0.1156	0.9049	5.6687
101.241	99.7764	0.2254	0.9477	0.1391	0.8927	5.3545
97.5001	102.9654	0.2566	0.9593	0.1243	0.8859	5.2993
101.1346	101.5677	0.2643	0.9554	0.1211	0.883	5.3099
107.6378	95.1343	0.2109	0.9699	0.1213	0.9004	5.3964
113.9639	95.9215	0.2627	0.9593	0.1065	0.8804	5.3521

For classification, a total number of 180 mammogram images which consist of 60 normal, 60 benign and 60 malignant were used. To begin classification, it is important to train the classifier. Here, two classes were defined to train the SVM algorithm in each of the phase of the classification technique applied. In the first phase of the classification, the first class has a unique label, +1 for cancerous mammograms while the second class has a unique label, -1 for non-cancerous mammograms, and in the second phase of the classification, the first class has a unique label, +1 for malignant mammogram images while the second class was assigned a unique label, -1 for benign mammogram images. These were used together with the extracted features for classification. Using the SVM classifier in the two phases, a training was carried out with 60% of the dataset (i.e., 36 images from each class of data) as well as testing with 40% of the dataset (i.e., 24 images from each class of data). The results of the classification and/or the performance of the two-phase SVM classifier are as shown in table 2 below:

TABLE II. CLASSIFICATION RESULTS

	Classification stage 1: Cancerous /noncancerous (out of 180 images)	Classification stage 2: Malignant/Benign (out of 120)
TP	173	116
FN	7	4
TN	175	118
FP	5	2

For classification stage 1;

Specificity=(TN/(TN+FP))*100=(175/(175+5))*100=97.22%

Sensitivity=(TP/(TP+FN))*100=(173/(173+7))*100=96.11%

Classification Rate (Accuracy)=

((TP+TN)/(TP+TN+FP+FN))*100=

((173+175)/(173+175+5+7))*100= 96.67%

Where TP (True Positive) = cancerous mammograms classified as cancerous mammograms;

FN (False Negative) = cancerous mammograms classified as noncancerous mammograms;

TN (True Negative) = Noncancerous mammograms classified as noncancerous mammograms;

FP (False Positive) = Noncancerous mammograms classified as cancerous mammograms.

For classification stage 2;

Specificity=(TN/(TN+FP))*100=(118/(118+2))*100=98.33%

Sensitivity=(TP/(TP+FN))*100=(116/(116+4))*100=96.67%

Classification Rate (Accuracy)=

((TP+TN)/(TP+TN+FP+FN))*100=

((116+118)/(116+118+2+4))*100=97.5%

Where TP (True Positive) = Malignant classified as malignant; FN (False Negative) = Malignant classified as benign;

TN (True Negative) = Benign classified as benign; FP (False Positive) = Benign classified as malignant. Figure 11 shows a graph illustrating the results of the SVM two-phase classification.

IV. CONCLUSION

The emphasis of this study is on false detection of breast cancer with the use of mammographic images and how the rate of false detection can be minimized. An automated technique for a two-phase SVM classification of breast cancer by the use of the combination of GLCM and SVM has been presented for this purpose. Besides the acquisition of mammogram images from MIAS database, the methodology consists of four main stages including image preprocessing, segmentation, features extraction and classification. These four stages of computation were carried out in MATLAB software (version 8.1). The first phase of the SVM classification identifies the mammogram images as either cancerous or non-cancerous while the second phase of the SVM classification identifies the cancerous mammogram images as either malignant or benign. From the graph illustrating the results of the SVM classification in figure 11, it is obvious that the rate of false detection by this technique has been greatly reduced. From the performance evaluation, the total accuracy of the classifier in the final phase of the classification is 97.5%. Considering this performance, it can be concluded that Computer Aided Diagnosis system put forward here can be used by doctors or radiologists to accurately identify breast cancer in the early stage.

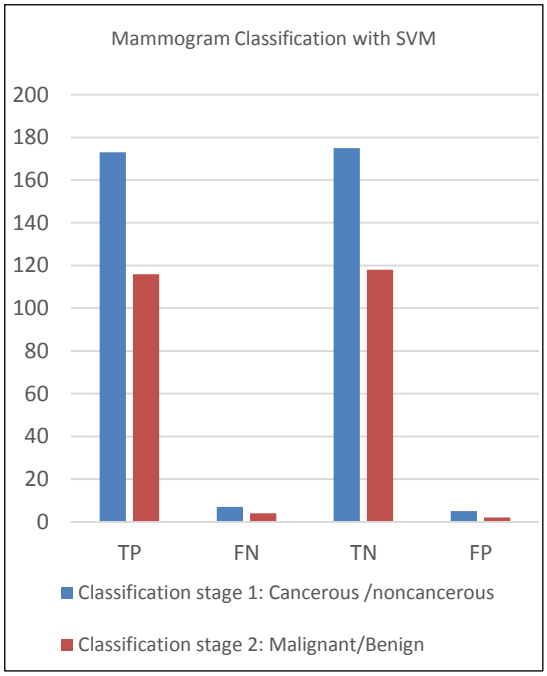

Figure 11. Graph illustrating the Results shown in Table 2.

REFERENCES

[1] Beheshti S. M. A., Noubari H. A., Fatemizadeh E. and Khalili M. (2016). "Classification of abnormalities in mammograms by new asymmetric fractal features", Biocybern Biomed Eng.; 36(1): 56-65. DOI: 10.1016/j.bbe.2015.07.002.

[2] Chatzistergos S. E., Andreadis I. and Nikita K. S. (2018). "Identification of architectural distortions in mammograms using local binary patterns and radial lengths through an exhaustive evaluation framework", Expert Syst.; 35(4): e12281. DOI:10.1111/exsy. 12281.

[3] Deepa S. N. and Aruna Devi B. (2011) et.al "A survey on artificial intelligence approaches for medical image classification", Indian Journal of Science and Technology, Vol. 4 No. 11, pp. 1583-1594.

[4] Ireaneus Anna Rejani Y. and Thamarai Selvi S. (2009). "Early Detection of Breast Cancer Using SVM Classifier Technique", International Journal on Computer Science and Engineering Vol. 1 (3), 127-130.

[5] Meenalosini, S., Janet J. and Kannan E. (2012). "A Novel Approach in Malignancy Detection of Computer Aided Diagnosis", American Journal of Applied Sciences 9 (7): 1020-1029, ISSN 1546-9239.

[6] Ragupathy U. S. and Saranya T. (2012). "Gabor Wavelet based Detection of Architectural Distortion and Mass in Mammographic Images and Classification using Adaptive Neuro Fuzzy Inference System" International Journal of Computer Applications, Vol. 46– No.22,pp. 0975 – 8887.

[7] Salve S. M. and Chakkarwar V. A. (2013). "Classification of Mammographic images using Gabor Wavelet and Discrete Wavelet Transform" *International Journal of Advanced Research in ECE ISSN: 2278-909X,Vol. 2* pp. 573-578.

[8] Setiawan A. S., Elysia, Wesley J. and Purnama Y. (2015). "Mammogram Classification using Law's Texture Energy Measure and Neural Networks". Procedia Comput Sci.; 59:92-7. DOI: 10.1016/j. procs.2015.07.341.

[9] Snehal A. M. and Kulhalli K. V. (2015). "Mammogram Image Features Extraction and Classification for Breast Cancer Detection", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume: 02, Issue: 07.

[10] Taheri Mohammad (2017). "Enhanced Breast Cancer Classification with Automatic Thresholding Using Support Vector Machine and Harris Corner Detection", Open Public Research Access Institutional Repository and Information Exchange (Open PRAIRIE).

How to Cite this Article:

Ikotun, A. M., Opiarighodare, D. K., Adelokun, A. P., Okikiola, F. M. & Lawal, O. N. (2020). Computer Aided Diagnosis of Breast Cancer in Human Mammogram Using Support Vector Machine. International Journal of Science and Engineering Investigations (IJSEI), 9(106), 44-51. http://www.ijsei.com/papers/ijsei-910620-07.pdf