International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Mammogram Images Classification using Linear Discriminant Analysis Technique

Ashraf Mohammed[1], Ali Ahmed[2], Waleed Mohammed[3], G.K.Viju[4], Mazin Taha[5]**

[1]Lecturer, Karary University, Khartoum, Sudan
[2]Associate Professor, King Abdulaziz University, Jeddah, Saudi Arabia
[3]Lecturer, Karary University, Khartoum, Sudan
[4]Professor, University of Garden City, Khartoum, Sudan
[5]Engineer, Republican Palace of Sudan Government, Khartoum, Sudan

---

Abstract -*Breast cancer represents a significant percentage of cancer death among women all over the world. Studies showed that breast cancer possibility of cure can increase up to 40%, if it detected still in early stage. The purpose of this paper is to use and apply a machine learning approach for mammogram image classification prediction in each image by classifying mammogram image to determine either it is benign breast tumors or malignant ones and help doctors to detect the disease in its early stage. This paper used Linear Discriminant Analysis (LDA classifier) and six statistical features that extracted from MIAS dataset, the dataset splits into two parts training and testing. After trained the classifier based on training set, the classifier tested based on the test set to determine the accuracy of the classifier depend on the confusion matrix. The paper emphasis of five phases starting by collecting images, preprocessing, features extracting, classification and end with testing and evaluating. The result of the proposed method empirically comes as 0.81% accuracy when using percentage of 85% of data set for training and 15% of data set for testing. Due to the increasing applications of ML methods in breast cancer research, we presented here an applied study of the Linear Discriminant Analysis method in the classification of mammogram images. Consequently, this study can assist in building computer aided diagnosis (CAD) systems in early detection of breasts Cancer .Consequently, we have contributed to this important field for biomedical research that may reduce the risk of late breast effects cancer.*

Key Words*: *Breast cancer, image processing, data mining, classification, linear discriminant analysis, Machine Learning, mammogram.

1. INTRODUCTION

Cancer is the uncontrolled growth of abnormal cells in the body [1].It is one of the leading causes of death which accounts for 13% of total deaths, worldwide .Breast cancer fall under main categories according to World Health Organization (WHO) cancer facts sheet 1 [2].Breast cancer is the most common cancer among women and one of the most important causes of death among them [3].Breast Cancer occurs when the cell tissues of breast become abnormal and divide uncontrollably. These abnormal cell tissues form large lumps, which consequently become a tumor [4].Studies showed that breast cancer possibility of cure can increase up to 40%, if it detected still in early stage. Mammography is considered the most useful technique for breast cancer early detection and it is still used all over the world [5].Indeed, early diagnosis of breast cancer significantly increases not only the number of treatment options available, but also the chance of success and survival of treatment [6, 7].For early detection, three methods are typically recommended and used in conjunction with each other – personal (self-made) breast exams, clinical (done by the doctor) breast exams, and mammograms [8].Mammography is one of the methods used for early detection of breast cancer by imaging a woman's breast. Then mammography helps doctors diagnose and classify images to determine whether there is a benign or malignant tumor.Breast Cancer has been a large topic in research area for the last few decades. A lot of research work has been done in the medical field to detect cancer. Still, the progress in detection and diagnosis of breast cancer remains very time to consume and expensive [9].Machine Learning is the most prominent sub-field of Artificial Intelligence that involved self-learning techniques that derived knowledge from data in order to make predictions. Machine Learning provides a more efficient way of capturing the knowledge in data to gradually improve the performance of predictive models and make data-driven decisions [10]. By development of data mining technology, it is not only extensively applied in commercial purposes, but also successfully applied in many different applications like medical tasks, for examples breast cancer screening. There are different machine learning techniques available for classification purpose that can differentiate breast cancer into benign (non-cancerous) or a malignant (cancerous). Therefore, there are many methods used for prediction and classification, such as: Decision Tree, Naive Bayes, Logistic Regression, k Nearest Neighbors and Discriminant Analysis. However, Machine Learning techniques have been applied in various domains and it has been proved that their practice is unavoidable in various applications [11]. In the present paper, using Linear Discriminant Analysis (LDA), an attempt is made towards correctly predicting the class of breast cancer (benign or malignant) and to help doctors diagnose

disease at an early stage to reduce the risk of fatal disease. The paper proceeds as follows. Section 2 reviews the study background and develops our study. In Section 3, we describe the proposed method. Our experimental and results is discussed in Section 4 and Section 5 presents the conclusion.

2. LITERATURE REVIEW

Diagnosing breast cancer is one of the most important stages in treating disease in the medical and health field. Numerous studies have been done by researchers to help doctors diagnose and detect the disease in its early stages. Machine learning is especially used in the diagnosis of breast cancer. In this part, we review some previous studies that dealt with this aspect through the application of many machine learning techniques. Study of Abhishek Midya and Jayasree Chakraborty (2015), entitled: Classification of benign and malignant masses in mammograms using multi-resolution analysis of oriented patterns. The study proposes a novel approach for the classification of breast masses as benign and malignant using multi-resolution analysis of oriented patterns of tissues in mammograms. Since, the oriented structures of normal breast near the mass region may be changed in presence of masses, three regions are defined, first, for the analysis. Statistical features are then extracted using two angle co-occurrence matrices derived at different resolution levels of each region with Haar-wavelet transform to quantify the joint occurrences of different angle pairs of oriented patterns. The experiments show best classification accuracy of 81.23% and area under the receiver operating characteristic curve of 0.86 with 433 images from the DDSM database using artificial neural network and tenfold cross-validation method [12]. Study of Zhicheng Jiao, Xinbo Gao, Ying Wang and Jie Li (2016), entitled: A deep feature based framework for breast masses classification. In this study, they design a deep feature based framework for breast mass classification task. It mainly contains a convolutional neural network (CNN) and a decision mechanism. Combining intensity information and deep features automatically extracted by the trained CNN from the original image, the proposed method could better simulate the diagnostic procedure operated by doctors and achieved state-of-art performance. In this framework, doctors′ global and local impressions left by mass images were represented by deep features extracted from two different layers called high-level and middle-level features. Meanwhile, the original images were regarded as detailed descriptions of the breast mass [13]. Study of Saima Anwar Lashari, Rosziati Ibrahim, Norhalina Senan, Iwan Tri Riyadi Yanto and Tutut Herawan (2016), entitled: Application of Wavelet De-noising Filters in Mammogram Images Classification Using Fuzzy Soft Set. This study proposed a classifier based on fuzzy soft set with embedding wavelet de-noising filters. Therefore, the proposed methodology involved five steps namely: MIAS dataset, wavelet de-noising filters hard and soft threshold, region of interest identification, feature extraction and classification. Experimental results show that proposed classifier FussCyier provides the classification performance with Daub3 (Level 1) with accuracy 75.64% (hard threshold), precision 46.11%, recall 84.67%, F-Micro 60% [14]. Study of Yu-Dong Zhang, Chichun Pan, Xianqing Chen, Fubin Wang (2018), entitled: Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. They proposed an improved nine-layer convolutional neural network (CNN). In addition, we compared three activation functions: rectified linear unit (ReLU), leaky ReLU, and parametric ReLU. Besides, six pooling techniques were compared: average pooling, max pooling, stochastic pooling, rank-based average pooling, rank-based weighted pooling, and rank-based stochastic pooling. The results over 100 test set showed the combination of parametric ReLU and rank-based stochastic pooling performed the best, with sensitivity of 93.4%, specificity of 94.6%, precision of 94.5%, and accuracy of 94.0% [15]. Study of Ashok Kumar, Saurabh Mukherjee and Ashish Kr. Luhach (2019), entitled: Deep learning with perspective modeling for early detection of malignancy in mammograms. Objective of this study is to deliver a classification system that can be used to classify breast images as a benign or malignant and if malignant then can further classify which type of malignancy that is non-invasive or invasive cancer. This model can also prescribe treatment for predicted malignant class with details like time taken, degree of seriousness, probability of curing by opted treatment because treatment of a breast cancer depends on type and stage of malignancy. To achieve higher or clinical usage accuracy by deploying advances of soft computing and image analysis like deep learning and deep neural networks to decrease breast cancer death as a concrete effort using mammograms by detecting breast cancer in an early stage[16].

3. PROPOSED METHOD

The main objective of the present study is to develop a machine learning approach for mammogram image classification prediction in each image and help doctors to detect the Breast cancer disease in its early stage. The proposed system consists of several phases as showed in Figure 1 followed by details about each phase:

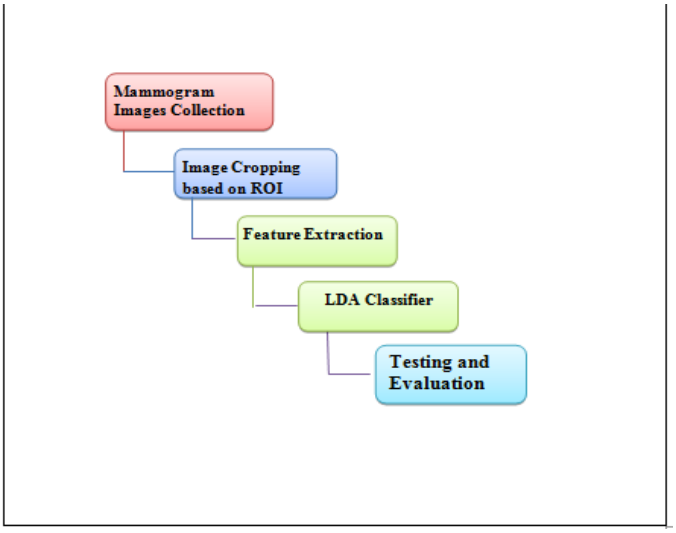

Fig -3.1: System Phases

3.1 Phase 1 (Mammogram Images Collection)

The Mammography Image Analysis Society (MIAS), which is an organization of UK research groups interested in the understanding of mammograms, has produced a digital mammography database [17].

3.2 Phase 2 (Image Cropping based on ROI)

A region of interest (ROI) is a portion of an image that you want to filter or perform some other operation on.

3.3 Phase 3 (Feature Extraction)

In this phase, after cropping the Region of Interest (ROI) from [x] position to [y] position and [radius] depend on the MIAS dataset. This stage applies the six functions to extract the features values from each mammogram images. The following paragraphs give more details about the six functions used to extract features values**.**

3.3.1 Mean

The Mean is a measure of the average intensity of the neighboring pixels of an image.

$$Mean = \sum_{i=0}^{L-1} Z_i * P(Z_i)$$

Equa -1: Mean

3.3.2 Standard Deviation

The Standard Deviation is a measure of how spreads out numbers are.

$$std = \sum_{i=0}^{l-1} (z_i - m)^2 * p(z_i)$$

Equa -2: Standard Deviation

3.3.3 Skewness

The Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. The Skewness for a normal distribution is zero, and any symmetric data should have Skewness near zero.Negative values for the Skewness indicate data that are skewed left and positive values for the Skewness indicate data that are skewed right.

$$skewness = \sum_{i=0}^{l-1} (z_i - m)^3 * p(z_i)$$

Equa -3: Skewness

3.3.4 Kurtosis

The Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean.

$$kurtosis = \sum_{i=0}^{l-1} (z_i - m)^4 * p(z_i)$$

Equa -4: Kurtosis

3.3.5 Contrast

The Contrast is the difference in luminance and/or color that makes an object (or its representation in an image or display) distinguishable. In visual perception of the real world, contrast is determined by the difference in the color and brightness of the object and other objects within the same field of view.

$$contrast = \sum_{i=0}^{l-1} \sqrt{(z_i - m)^2 * p(z_i)}$$

Equa -5: Contrast

3.3.6 Smoothness

Measures the relative intensity variations in a region

$$smoothness = 1 - \frac{1}{(1 + \sigma^2)}$$

Equa -6: Smoothness

3.4 Phase 4 (LDA classification)

Applying Linear Discriminant Analysis classifier and calculate the classification accuracy.

3.5 Phase 5 (Testing and Evaluation)

Testing is an important part of the classification process. Evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, called confusion matrix.

3.6 Linear Discriminant Analysis (LDA)

There are many possible techniques for classification of data. Linear Discriminant Analysis (LDA) is commonly used technique for data classification and dimensionality reduction. Linear Discriminant Analysis easily handles the case where the within-class frequencies are unequal and their performances have been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal

separability[18]. Despite its simplicity, LDA often produces robust, decent, and interpretable classification results. When tackling real-world classification problems, LDA is often the first and benchmarking method before other more complicated and flexible ones are employed [19]. Figure 2 show the LDA classifer.

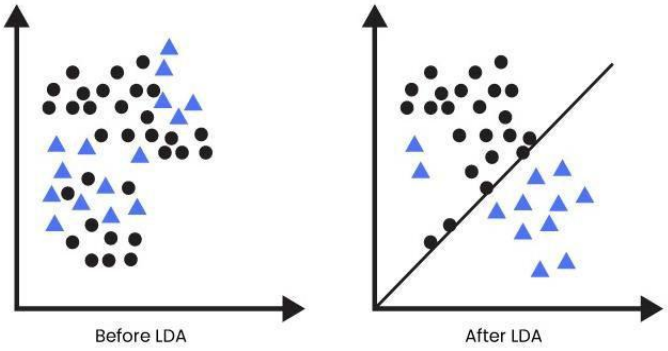

Fig -3.1: LDA classifer

3.7 Real-Life Applications of LDA

Some of the practical applications of LDA are listed below:

3.7.1 Face Recognition

LDA is used in face recognition to reduce the number of attributes to a more manageable number before the actual classification. The dimensions that are generated are a linear combination of pixels that forms a template. These are called Fisher's faces.

3.7.2 Medical

You can use LDA to classify the patient disease as mild, moderate or severe. The classification is done upon the various parameters of the patient and his medical trajectory.

3.7.3 Customer Identification

You can obtain the features of customers by performing a simple question and answer survey. LDA helps in identifying and selecting which describes the properties of a group of customers who are most likely to buy a particular item in a shopping mall [20].

3.8 Dataset

The data sets used for the study taken from Mammographic Image Analysis Society MIAS every image is 1024 × 1024 pixels. This database contains Benign and Malignant breast images for a total of 119 patients. The images also include the locations of any abnormalities that may be present [21].

3.9 Evaluation Measures

The performance of the classifier is estimated by using confusion matrix is a table that is often used to describe the performance of a classification model or "classifier" on a set of test data for which the true values are known. The structure of confusion matrix is presented in table 2 given below:

Table -1: Format of Confusion Matrix

True Class	Predicted Class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

confusion matrix, running time and classification accuracy. A The brief description of TP, TN, FP, and FN is given below:
True Negative (TN): No Possibility of disease, Prediction is false.
True Positive (TP): Possibility of disease, Prediction is true.
False Positive (Type 1 error): They do not have the disease but the prediction is true.
False Negative (Type 2 error): They have the disease, Prediction is true.

A number of different measures are commonly used to evaluate the performance of the proposed method. These measures including Accuracy, sensitivity and specificity calculated from confusion matrix using the following equations:

Accuracy = (TP+TN)/ (TP+TN+FP+FN), (Number of correct assessments)/Number of all assessments).

$$CR=\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Equa -7: Accuracy

Specificity= TN/ (TN + FP) = (Number of true negative assessment)/ (Number of all negative assessment)

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

Equa -8: Specificity

Sensitivity: is the percentage of positive records classified correctly out of all positive records.
Recall= TP/ (FP+TN), (Number of true positive assessment)/ (Number of all positive assessment)

$$\text{Sensitivity}=\frac{TP}{TP+FN}$$

Equa -9: Sensitivity

By applying the steps mentioned in this part of the paper, various results of the classification process can be obtained through the linear discrimination technique in the next part of the paper.

4. EXPERIMENTAL AND RESULTS

The classification process is divided into training (known data are given to the technique for training) and testing part (unknown data are given to the technique). In this study, the experiment was done according to the following sequence, firstly collecting mamogram images, secondly cropping each image based on (X, Y) and R to detecting the region of interest, thirdly extracting the six statistical features, finally applying LDA classification algorithm using the statistical features (the experiment was done three times) and calculate the accuracy of classification process. The tables and figures below describes classification process and accuracy for the three different sizes of training and testing dataset.

Table - 4.1: shows the resulting classification process for the size of the data set (60-40), (real positive rate, recall, and accuracy).

dataset size	Experiment no	TP rate	Recall	Accuracy
60-40	A	0.4167	0.2778	0.5238
	B	0.5000	0.4444	0.5714
	C	0.3333	0.0556	0.5476

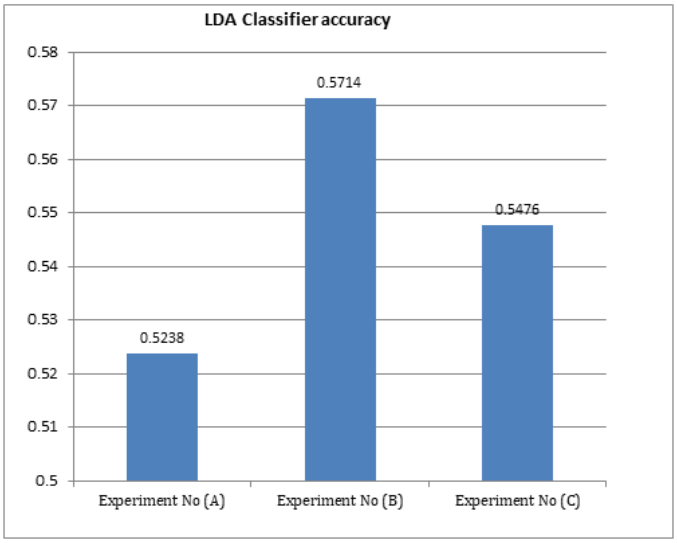

Chart - 4.1: LDA Classifier accuracy for the size of dataset (60-40)

Table -4.2: shows the resulting classification process for the size of dataset (70-30)

dataset size	Experiment no	TP rate	Recall	Accuracy
70-30	A	0.6000	0.2143	0.5938
	B	0.7143	0.3571	0.6563
	C	1	0.1429	0.6250

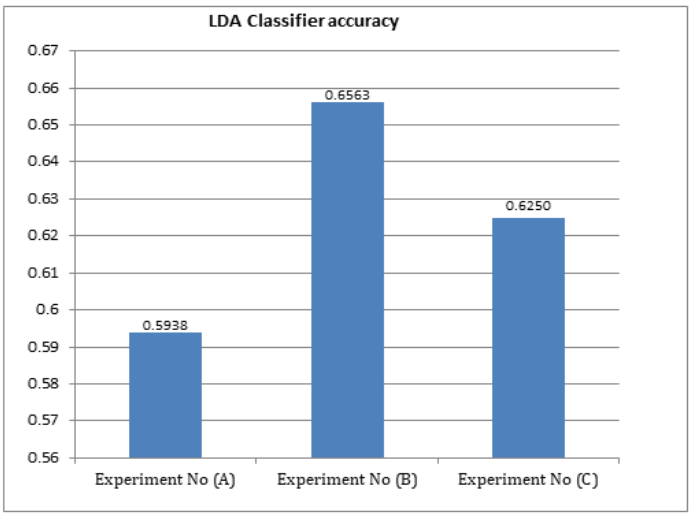

Chart - 4.2: LDA Classifier accuracy for the size of dataset (70-30)

Table - 4.3: shows the resulting classification process for the size of dataset (85-15)

dataset size	Experiment no	TP rate	Recall	Accuracy
85-15	A	1	0.2857	0.6875
	B	1	0.5714	0.8125
	C	1	0.4286	0.7500

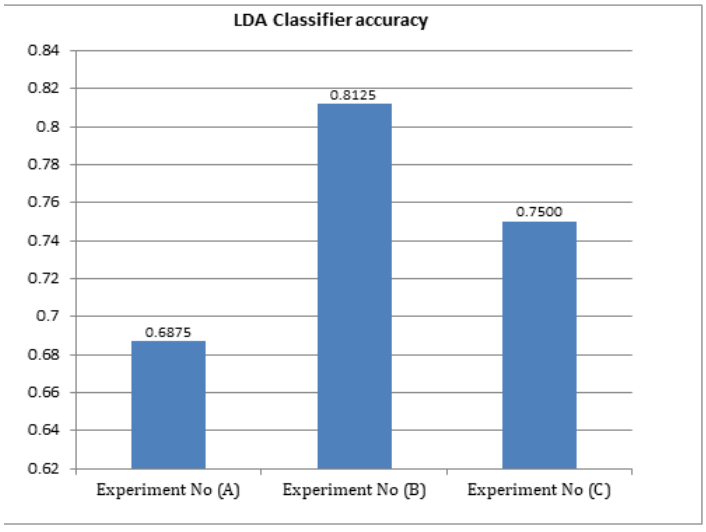

Chart - 4.3: LDA Classifier accuracy for the size of dataset (85-15)

Different three sizes of training data was chosed randomly, these sizes are 60, 70 and 85% percentages of data, respectively.of each size training dataset and testing dataset, we compared accuracy for the different three sizes of dataset (60-40), (70-30) and (85-

15).We found out that the highest classification accuracy result is obtained when the size of training are chosen is (85%).Table 4.4 shows the highest classifications results Accuracy for different dataset sizes.

Table - 4.4: shows the highest classifications results Accuracy for different dataset sizes.

dataset size	Experiment no	TP rate	Recall	Accuracy
60-40	A	0.5000	0.4444	0.5714
70-30	B	0.7143	0.3571	0.6563
85-15	C	1	0.5714	0.8125

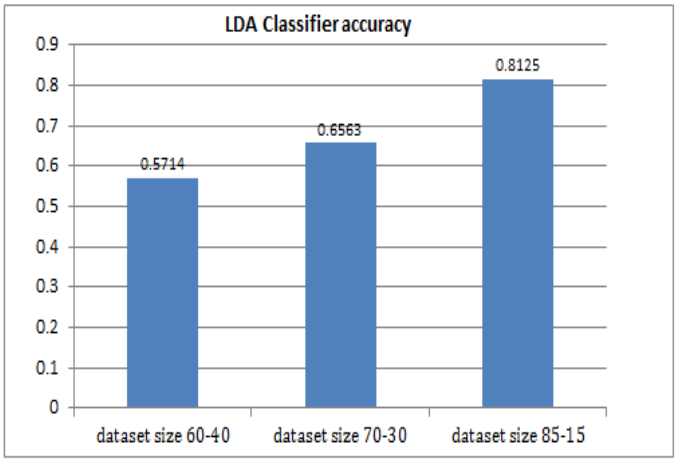

Chart - 4.4: The highest classifications results Accuracy for different dataset sizes

5. CONCLUSIONS

Breast cancer is one of the main reason for death in the female. So early detection and diagnosis of breast cancer are very important in reducing life losses. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones. This study applied LDA classifier using six statistical features extracted from each mammogram image downloaded from MIAS data set. The accuracy of classification prcess depends on two things: the accurate features and the classifier method used. Accurate features play an important role in classification accuracy. The best accuracy is (81%) obtained when split the data in to (85%) for the training and (15%) for testing. Consequently, this study can assist in building computer aided diagnosis (CAD) systems in early detection of breasts Cancer that may reduce the risk of late breast cancer effects.

ACKNOWLEDGEMENT

We would like to take the opportunity to express our heartiest gratitude towards all those people who have, in various ways helped us to complete the research. We thank Dr. Faisal Mohammed, Associate Professor and Dr. Ibrahim Mohammed, Associate Professor for their helpful comments. The paper has also benefited from comments by participants at the scientific seminar of the University of Karary in 2019. We also thank the family of Al-Neelain University represented in the College of Computer Science and Information Technology for its support in providing us with the necessary aids to extract this paper.

REFERENCES

[1] Cancer treatment center of America. (2020). what is cancer? Retrieved from https://www.cancercenter.com/what-is-cancer.

[2] Rao, P. B., & Deeba, F. (2020). Expressions of biomarkers in MCF7 Breast and Colon Cancer Cell Lines. *Journal of Drug Delivery and Therapeutics*, *10*(2), 107-114.

[3] Momenimovahed, Z., & Salehiniya, H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets and Therapy*, *11*, 151.

[4] Mehdy, M. M., Ng, P. Y., Shair, E. F., Saleh, N. I., & Gomes, C. (2017). Artificial neural networks in image processing for early detection of breast cancer. Computational and mathematical methods in medicine, 2017.

[5] Oliver, A., Lladó, X., Pérez, E., Pont, J., Denton, E. R., Freixenet, J., & Martí, J. (2010). A statistical approach for breast density segmentation. *Journal of digital imaging*, *23*(5), 527-537.

[6] Laronga, C., Chagpar, A. B., & Vora, S. R. (2016). Patient education: breast cancer guide to diagnosis and treatment (beyond the basics). *UpToDate. UpToDate, Waltham. https://www. uptodate. com/contents/breast-\ cancer-guide-to-diagnosis-and-treatment-beyond-the-basics. Accessed*, *7*.

[7] Scharl, A., Kühn, T., Papathemelis, T., & Salterberg, A. (2015). The right treatment for the right patient–personalised treatment of breast cancer. *Geburtshilfe und Frauenheilkunde*, *75*(07), 683-691.

[8] Fuller, M. S., Lee, C. I., & Elmore, J. G. (2015). Breast cancer screening: an evidence-based update. *The Medical clinics of North America*, *99*(3), 451.

[9] Giri P, Saravanakumar K (2017). Breast Cancer Detection using Image Processing Techniques. Orient. J. Comp. Sci. and Technol; 10(2). Available from: http://www.computerscijournal.org/?p=5299

[10] Raschka Sebastian and Mirjalili (2017),"Python Machine Learning", 2nd edition, Packt Publishing Ltd., ISBN: 978-1-78712-593-3.

[11] Nithya B.(2016),"An Analysis on Applications of Machine Learning Tools, Techniques an Practices in Health Care System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 6,Page(s)-1-8.

[12] Midya, A., & Chakraborty, J. (2015, April). Classification of benign and malignant masses in mammograms using multi-resolution analysis of oriented patterns. In 2015

IEEE 12th International Symposium on Biomedical Imaging (ISBI) (pp. 411-414). IEEE.

[13] Jiao, Z., Gao, X., Wang, Y., & Li, J. (2016). A deep feature based framework for breast masses classification. Neurocomputing, 197, 221-231.

[14] Lashari, S. A., Ibrahim, R., Senan, N., Yanto, I. T. R., & Herawan, T. (2016, August). Application of wavelet de-noising filters in mammogram images classification using fuzzy soft set. In International Conference on Soft Computing and Data Mining (pp. 529-537). Springer, Cham.

[15] Zhang, Y. D., Pan, C., Chen, X., & Wang, F. (2018). Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. Journal of computational science, 27, 57-68.

[16] Kumar, A., Mukherjee, S., & Luhach, A. K. (2019). Deep learning with perspective modeling for early detection of malignancy in mammograms. Journal of Discrete Mathematical Sciences and Cryptography, 22(4), 627-643

[17] Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., & Taylor, P. (2015). Mammographic image analysis society (MIAS) database v1. 21.

[18] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing, 18, 1-8.

[19] YANG Xiaozhou. Linear Discriminant Analysis, Explained. Retrieved from https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b.

[20] Priyankur Sarkar.Linear Discriminant Analysis for Machine Learning.(2019). Retrieved from https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning.

[21] Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., ... & Taylor, P. (2015). Retrieved fro Mammographic image analysis society (MIAS) database v1. 21.