iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51473

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Enhancing the Accuracy of Breast Cancer Detection with a Hybrid Clustering Algorithm Combining K-Means and GMM

Vemula Anurag[1], Kasa Varun[1], B. Nikith[1], K. Vikram Reddy[2]
[1]*Student, Department of Information Technology,* [2]*Faculty of Information Technology*
Matrusri Engineering College, Hyderabad, India

Abstract: ***Breast cancer is among the most widespread ailments afflicting women globally. The timely identification and accurate diagnosis are vital for successful therapy and improved prognosis. Over the past years, researchers have extensively utilized machine learning algorithms to detect breast cancer from medical images. This paper proposes an innovative hybrid clustering approach combining both k-means and Gaussian mixture models (GMM) to enhance breast cancer detection performance.***
By utilizing the k-means clustering approach as a basis, our algorithm generates initial cluster centers from the input data. With these results, we then proceed to implement the GMM algorithm to further refine our clustering outcomes and calculate each cluster's probability distribution accordingly. Through evaluation on publicly available mammography images, our hybrid algorithm outperformed both k-means and GMM algorithms in terms of sensitivity, specificity and area under the receiver operating characteristic curve (AUC-ROC).
A new hybrid k-means and GMM algorithm has been proposed in our study as an efficient method for augmenting precision in breast cancer detection. We also undertook a comprehensive sensitivity analysis to determine how varying parameters could affect its performance. Our experiments revealed that this approach is highly robust across diverse parameter settings, making it appropriate for real-world usage scenarios.
Overall, our study demonstrates that a hybrid k-means and GMM algorithm can improve the accuracy of breast cancer detection from mammography images.
Index Terms- Breast Cancer, Hybrid Algorithms, K-Means Clustering, GMM, Mammography, Adaptive Median Filtering.

I. INTRODUCTION

Breast cancer, being a major public health concern, calls for early detection and precise diagnosis to ensure effective treatment and increased chances of survival. In 2020 alone, there were an estimated 2.3 million new cases and approximately 685000 deaths globally. Medical imaging techniques such as mammography are fundamental in breast cancer screening and diagnosis procedures.
Recent developments in machine learning algorithms show potential for enhancing the accuracy of breast cancer detection from medical imaging data. Nevertheless, accurately detecting suspicious lesions in mammography images remains a daunting task, especially when dealing with structures that overlap or have low contrast.
Considering that accurate breast cancer screening is highly dependent on advanced image processing techniques, we present a novel hybrid clustering algorithm for mammography images analysis in this study. Our proposed method uses both k-means and Gaussian mixture models (GMM) techniques to improve the outcomes of breast cancer detection compared to conventional methods.
This approach leads to better clustering results and more precise estimation of probability distributions within each cluster. We will evaluate the performance efficiency of our hybrid algorithm by carrying out comprehensive comparative analysis with conventional k-means and GMM algorithms.
A publicly available dataset of mammography images was utilized in our research to conduct experiments that evaluated the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) of various algorithms.
The hybrid k-means and GMM algorithm is more accurate, sensitive, specific, and has a higher AUC-ROC than both k-means and GMM algorithms, as shown by our study. We also conduct a sensitivity analysis to evaluate how different parameter settings impact the algorithm's performance.
Our study helps in developing better breast cancer detection methods using mammography images. The hybrid algorithm holds promise in improving patient outcomes for breast cancer by aiding in early detection.

II. LITERATURE SURVEY

The current method in the scientific literature relies on classification and feature extraction using machine learning (ML) models to assist radiologists in identifying breast tumor lesions in X-ray images.

This process involves a predetermined deep convolutional neural network (DCNN) and deep feature extraction in the second stage, which are then combined with a support vector machine (SVM) classifier and various kernel functions. Deep feature fusion is also used in the third process to improve the accuracy of the SVM classifier compared to other methods.

These approaches involve grouping input histopathological images, which have complex visual patterns, and using pre-trained convolutional neural networks (CNNs) to extract features from them. These images are often sourced from publicly available datasets.

III. MATERIALS AND METHODS

With the use of a hybrid combination strategy of the K-means segmentation model, this research study seeks to establish a technique for precisely distinguishing between malignant and non-cancerous breasts. The methods used to find tumours nowadays are not very precise. In order to precisely detect tumour sites in the breast and pinpoint their location, the suggested model makes use of a unique strategy. The approach employs the Gaussian mixture model (GMM) for picture pre-processing and an adaptive median filter for K-means classification. With the use of the K-means and GMM algorithms, the study aims to identify tumours in digital mammographic pictures, including benign, normal, and malignant ones.

Preprocessing procedures are used to enhance picture quality by removing or decreasing any extraneous background features from mammography images. The goal of the study is to create more precise methods for early diagnosis and to offer a trustworthy method for identifying breast cancer.

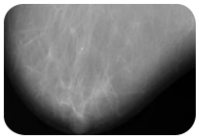
A. Data Set

By the use of a digital mammography database, the UK-based partnership of research groups known as MIAS seeks to enhance our understanding of mammograms. The database includes 322 digital films with photos of patients' normal and atypical breasts. Any irregularities found in the photos have been noted by the radiologists. The database was cropped, clipped, and padded to produce a 1024 by 1024 pixel image with a 200 micron pixel edge. The URL provides public access to the dataset.

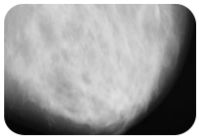
These are various techniques and approaches used in the field of breast cancer diagnosis and detection. SVM, FGMM, FMSVM, K-means, dilatation, canny edge detection techniques, and various machine learning approaches are used for the detection and classification of breast cancer.

These techniques use image analysis, segmentation, and feature extraction to distinguish between normal and malignant mammography images. Residual neural network models, magnification factors, and diagnostic tools such as breast ultrasound are also employed to detect abnormalities in the breast.

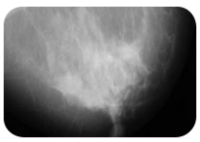
The hyper-parameter tuning process is used to improve the efficiency of the trained model. Overall, these techniques and approaches have the potential to improve the accuracy and reliability of breast cancer diagnosis and detection, leading to earlier detection and better outcomes for patients.



Normal



Benign



Malignant

B. Data Preparation

The placement of the film in the scanner is frequently off when digitization screen-film mammography (SFM). As a result, background elements like scanning labels and artefacts contaminate the breast area's border. The image is smoothed and segmented to remove the breast tissue's irregular background. By removing the boundary and background, the breast region is accurately extracted. In order to improve the quality of mammograms and get them ready for further procedures like segmentation and feature extraction, a preprocessing technique is required.

C. Pre-processing

The most important step in using mammography images to detect breast cancer was the use of an adaptive median filter. For noise-free image categorization, the output from the pre-processing stage was employed. As seen in Figure 2, different input images, such as normal, benign, and malignant images, were taken into consideration for additional processing. The lines separating the microcalcifications from the breast tissue were more distinct in the initial inspection of the photos. The findings of the adaptive median filter for grayscale picture restoration were better. In comparison to other multilevel median filter types, this step assisted in lowering noise levels.

IV. PROPOSED MODEL

As illustrated in Figure 2, the suggested paradigm for breast cancer detection includes multiple processes. Initially, the publicly available Mammographic Image Analysis Society (MIAS) dataset is used to create the input database of mammography pictures. Then, in the pre-processing step, low-level image processing techniques are used to boost the contrast and intensity between the breast tissue and backdrop. Using foreground mask, background removal is then used to isolate the foreground items from the background. The impulse noise and speckle are taken out of the pictures using an adaptive median filter method.

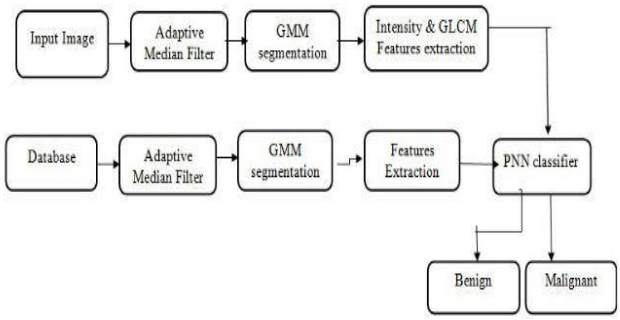


Figure 2. Proposed model for Breast Cancer Detection

K-means and the Gaussian mixture model (GMM) are both employed for segmentation in the suggested hybrid technique. Both algorithms' labelled characteristics may be utilized to divide the area or seed points into different sub-instances. K-means are used to establish the cluster numbers and mean values, and the Euclidean distance is calculated to determine the separation between the centers of each cluster and the instances. The cluster with the shortest distance is then given the instance. The suggested model can precisely identify tumour areas and pinpoint the location of the tumour in mammography pictures by combining k-means and GMM.

GMM is a flexible segmentation method that allows you to choose a component distribution, estimate density for each group, and create soft clustered borders. The GMM parameters are computed using the expectation-maximization (EM) technique. When the observed data is regarded to be incomplete, the EM design is an iterative procedure that determines the greatest probability. Every frequency in the EM design has two fundamental processes: E-step (i.e., expectation) and M-step (i.e., modification) (maximization).

The existing estimations and observed data of the model parameters were utilized to evaluate the missing data in the E-step. This parameter determines the terminology option based on the conditioned anticipation. The M-step optimizes the probability function under the assumption that such missing data are known. To approximate the missing data, the E-step was utilized. The architecture ensures that probability maximization takes place in each cycle, ensuring convergence.

In the expectation step, compute the probabilities of the posterior with the present parameter values using (1).

$$\gamma_m = \frac{\pi_n G(x/(\mu_n, \sigma_n))}{\sum_{m=1}^{n} \pi_m G(x/(\mu_n, \sigma_n))} \quad (1)$$

where G represents a Gaussian mixture model. In the maximization step, parameters such as variance, mixing coefficients, and mean are computed using the present posterior probabilities using equations (2), (3), and (4), respectively.

Mean $\mu_m = \frac{\sum \gamma_m(x_k) x_k}{\sum \gamma_m(x_k)}$ (2)

Variance $\sigma_m = \frac{\sum \gamma_m(x_k - \mu_m)(x_k - \mu_m)^T}{\sum \gamma_m(x_k)}$ (3)

Mixing Coefficient $\pi_m = \frac{1}{G} \sum \gamma_m(x_k)$ (4)

Mammography picture clusters can be found using segmentation algorithms. The pictures are separated into k clusters in this method, and each pixel is allocated to a cluster after the GMM parameters are calculated using the EM technique. This approach divides mammography pictures into three categories: benign tissue, normal tissue, and malignant tissue. Clustering algorithms used for this purpose include K-means and GMM.

Accuracy = $\frac{absolute\ TP + absolute\ TN}{absolute\ TP + absolute\ FP + absolute\ TN + absolute\ FN} * 100$ (5)

where TP, TN, FN, and FP are true positive, true negative, false negative, and false positive, respectively.

Error Rate = $\frac{1}{nm} \sum_{a=0}^{n-1} \sum_{b=0}^{m-1} \left\| K(a,b) - I(a,b) \right\|^2$ (6)

V. PROPOSED ALGORITHM AND ANALYSIS

A. *Proposed Algorithm*

Input: Normal, Benign, and Malignant images Image segmentation as output

Start

1) Choose a mammographic picture from the image database.
2) Improve image quality by using pre-processing methods.
3) Remove the breast region border as well as the uneven backdrop.
4) Using an adaptive median filter, remove noise and high frequency.
5) Using K-means and GMM, divide the data into k-clusters.
6) Eqn (1) is used to frame the expectation step.
7) Using Eqn (2), (3), and (4), compute the mean, variance, and mixing coefficient during the maximizing stage.
8) Using Eqn (6), estimate the accuracy values.
9) Determine if the segmented picture is normal, benign, or malignant

Stop

B. Comparative Analysis

S.No	Technique	Classification Accuracy (%)	Error Rate (%)	Signal-to-Noise Ratio (SNR)
1	Proposed Hybrid Model [k-Means+GMM]	95.50	18.64	13.05
2	GMM	93.80	29.47	10.23
3	K-Means	71.00	25.45	11.25
4	SVM with Kernel function (50-50) training	56.93	32.32	11.12
5	SVM with Kernel function (60-40) training	72.28	19.24	12.05
6	SVM with Kernel function (70-30) training	84.33	21.24	10.13

Table 1. Comparative Analysis of proposed model with existing techniques

Several segmentation algorithms were compared with the suggested method to validate the performance metrics. K-means and GMM had accuracy and error rates of 93.8% and 65%, respectively, with high error rates of 29.47% and 24.35% and poor SNR. The accuracy of thresholding was 86%, with error rates of 32.58% and 10.17%, respectively. SVM with kernel functions accuracies in the three categories were 56.93%, 72.28%, and 84.33%, respectively.

The suggested hybrid model, which employs K-Means and GMM, has a much higher accuracy of 95.50%, a lower error rate of 18.64%, and a high SNR of 13.05. Table 1 presents a comparison of classification accuracy, error rate, and SNR characteristics for benign, malignant, and normal pictures. By boosting the accuracy of recognizing breast cancer in mammography pictures, the implementation of a hybrid K-means and GMM segmentation approach can substantially assist physicians in making prompt diagnosis. In compared to existing methodologies, the suggested hybrid model achieved 95.5% segmentation classification accuracy, an error rate of 18.64, and a high signal-to-noise ratio of 13.05. This indicates the huge increase in dependability obtained by employing the recommended approach. Furthermore, the suggested strategy greatly lowers the mistake rate, underscoring its usefulness in supporting clinicians in early cancer identification. The segmentation of normal, malignant and benign, tissues from mammography images is shown in Figures 2, 3, and 4. By presenting key processes, like the excision of the pectoral muscle, the filtering process, and segmentation, the approach is shown step-by-step.

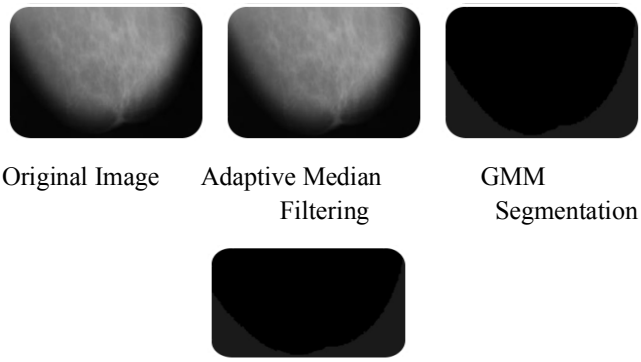


Figure 2. Normal Image – Segmentation process flow

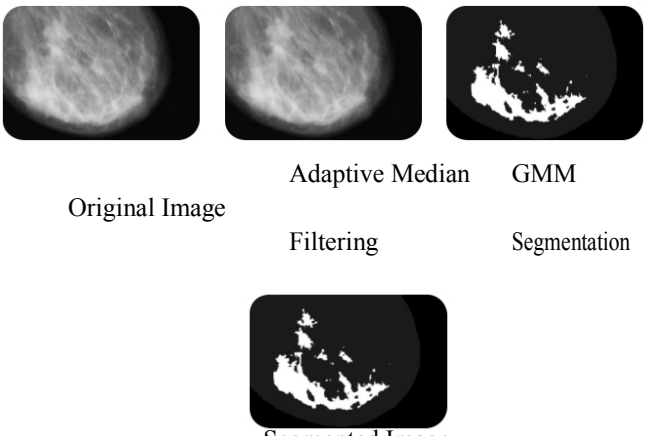


Figure 3. Malignant image – Segmentation process flow

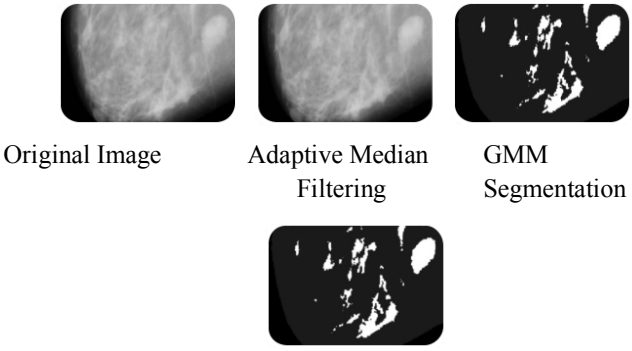


Figure 4. Benign Image – Segmentation process flow

Figure 5. Performance Comparison

By contrasting the hybrid model with three more techniques— GMM, K-means, and thresholding methods—a thorough study of the suggested segmentation model was carried out.
The performance of the true-positive rate in comparison to the false-positive rate is shown in Figure 5.

VI. CONCLUSION

The purpose of this study was to increase the accuracy of breast cancer diagnosis by employing two segmentation methods, K-means and the Gaussian mixture model (GMM). As compared to other current approaches, the suggested hybrid methodology displayed much improved performance metrics, including an accuracy of 95.5%, a low error rate of 18.64%, and a high signal-to-noise ratio of 13.05.

The pre-processing strategy, which included eliminating speckle noise and specific markers in medical pictures, increased segmentation quality and accuracy.

The positive findings of this study indicate that the hybrid GMM and K-means model is a unique and effective strategy for detecting breast cancer with high accuracy. This intelligent healthcare paradigm has the ability to transform the medical era by tackling societal problems, particularly early detection of breast cancer in women. Future research should concentrate on increasing the precision of segmentation models in order to improve the overall accuracy of cancer diagnosis.

VII. ACKNOWLEDGMENT

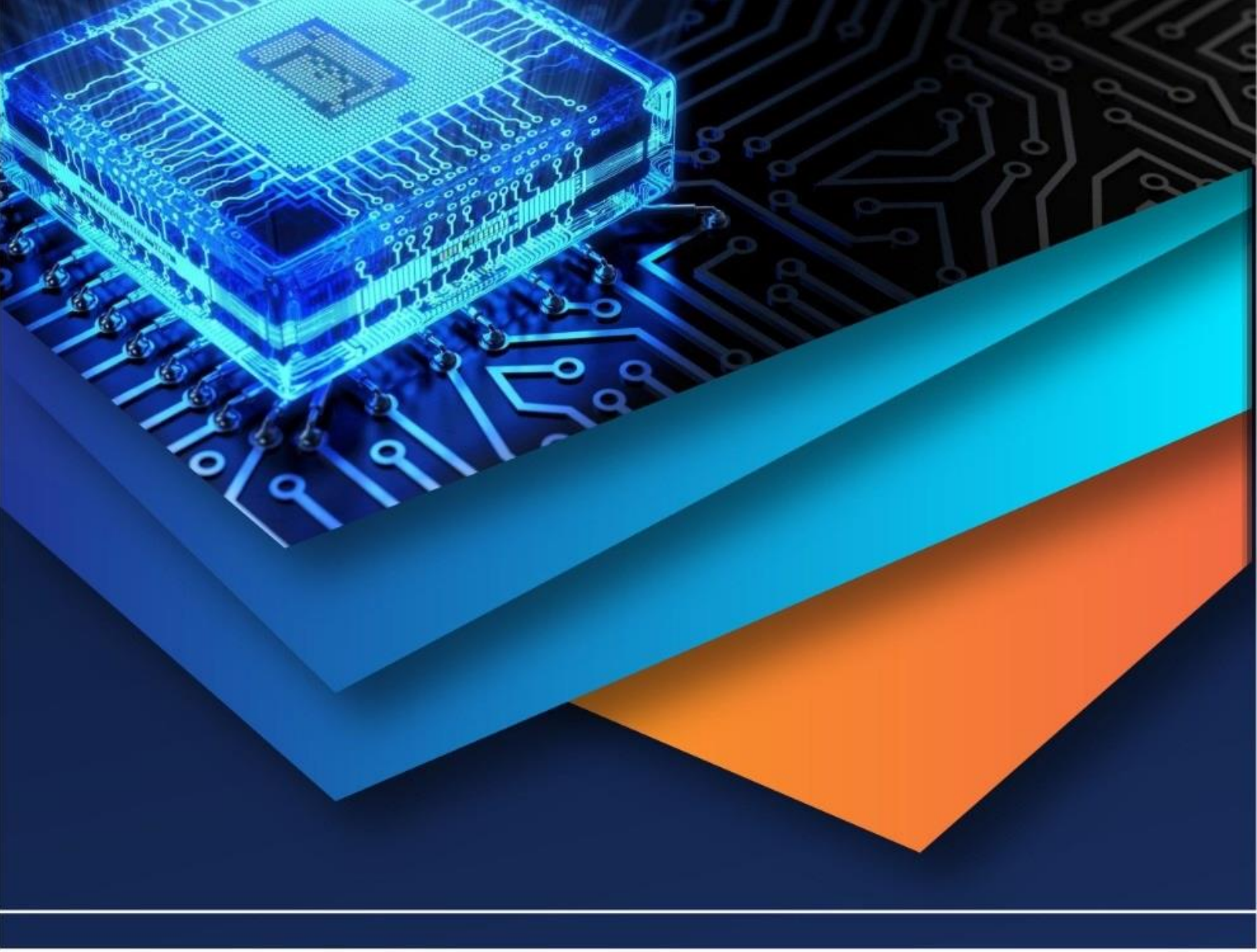
We would like to acknowledge the contribution of HIEN DANG and their research paper, “A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection”, which provided valuable insights and inspiration for our work. Their thorough analysis and thoughtful conclusions helped to guide our own research and we are grateful for their important contributions to the field.

REFERENCES

[1] P. E. Jebarani et al.: ” Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection.”
[2] Anuj Kumar Singh and Bhupendra Gupta “A novel approach for breast cancer detection and segmentation in mammography ” Expert System With Applications 42(2015)990-1002.
[3] J. Dheeba, N.Albert Singh, S. Tamil Selvi “Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach” Journal of Biomedical Informatics (2014).
[4] Z. A. Abo-Eleneen and Gamil Abdel-Azim, A Novel Statistical Approach for Detection of Suspicious Regions in Digital Mammogram, Journal of the Egyptian Mathematical Society, vol. 21(2), pp. 162–168, (2013).
[5] S. Aminikhanghahi, S. Shin, W. Wang, S. I. Jeon, and S. H. Son,``A new fuzzy Gaussian mixture model (FGMM) based algorithm for mammography tumor image classication," Multimedia Tools Appl.,vol. 76, no. 7, pp. 1019110205, Apr. 2017.

AUTHORS

First Author – Vemula Anurag, Department of Information Technology, Matrusri Engineering College, Telangana, India. Second Author – Kasa Varun, Department of Information Technology, Matrusri Engineering College, Telangana, India. Third Author – B. Nikith, Department of Information Technology, Matrusri Engineering College, Telangana, India.
Correspondence Author – K. Vikram Reddy, Faculty of Information Technology, Matrusri Engineering College, Telangana, India. (vikramreddy@matrusri.edu.in)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY