

Received October 12, 2021, accepted October 23, 2021, date of publication October 27, 2021, date of current version November 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123425

# A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection

P. ESTHER JEBARANI<sup>1</sup>, N. UMADEVI<sup>1</sup>, HIEN DANG<sup>1,2,3</sup> , (Member, IEEE), AND MARC POMPLUN<sup>3</sup>

<sup>1</sup>Department of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore 641005, India

<sup>2</sup>Faculty of Computer Science and Engineering, Thuyloi University, Hanoi 100000, Vietnam

<sup>3</sup>Department of Computer Science, University of Massachusetts Boston, Boston, MA 02125, USA

Corresponding author: Hien Dang (hiendt@tlu.edu.vn)


This work was supported in part by Thuyloi University under Grant NCKH2021.

**ABSTRACT** Breast cancer is the second leading cause of death among a large number of women worldwide. It may be challenging for radiologists to diagnose and treat breast cancer. Consequently, primary care improves disease prevention and death. Early detection increases treatment options and saves life, which is the major target of this research. This research indicates the versatility of the methodology by integrating contemporary segmentation approaches with machine learning methods, which are developing areas of research. In the pre-processing process, an adaptive median filter is utilized for noise removal, enhancement of image quality, conservation of edges, and smoothing. This research makes a significant contribution by proposing a new parameter for evaluating K-means and a Gaussian mixture model (GMM) performance. A hybrid combination of segmentation and detection was applied to breast cancer. The proposed technique is significant for classifying benign and malignant tumors. The simulated results are discussed and evaluated to determine the competence of this method for the early diagnosis of breast cancer. This method allows medical experts to recognize breast cancer at a faster rate and provide higher accuracy. An ANOVA test was used to determine the multi-variant analysis and prediction rate for the proposed method.

**INDEX TERMS** Breast cancer, pre-processing, adaptive median filtering, K-means, EM algorithm, Gaussian mixture model, ANOVA.

## I. INTRODUCTION

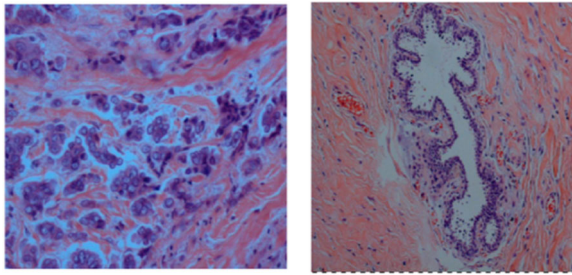
Experts in modern medical areas are focusing more on technical approaches for a variety of chronic diseases. Even though many diseases are incurable, such as cancer, stroke, heart attack, chronic liver diseases, viral hepatitis, and coronary artery disease, the death rate from breast cancer is increasing every year. According to a statistical report on medical health, cancer is a genetic disease that leads to variations in genes involved in the functionality of human body cells. Variation of the gene in genetic diseases may affect the internal parts of human organs for future generations. It may also affect DNA structure, resulting in environmental exposure to substances such as UV radiation, smoking, and other variables that are significant in the development of breast cancer [1]. Despite this, 60% of women affected by breast cancer are diagnosed at the last stage, which leads to death in women.

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano .

The main contribution of the proposed method is to segment the disordered portion of the cancerous cells in the breast image. The novel idea in this work includes a hybrid technique for determining breast cancer detection, and multi-variant analysis is performed to improve the prediction rate for the proposed system.

Research on breast cancer has increased in the past decade when abnormalities and uncontrollability in breast cell tissues develop into serious breast cancer in women [2]. It may include angiosarcoma, ductal carcinoma in situ (DCIS), and lobular carcinoma in situ (LCIS). As a result, it is critical to track the number of deaths caused by breast cancer before treatment. Figure 1 (a) and (b) show cancerous and non-cancerous images taken as exemplars. Therapeutic imaging is a non-invasive method of examining the inside of the human body that can help doctors detect and treat breast cancer at an early stage.

The determination of breast cancer in the initial stage is controllable. Breast cancer is caused by microcalcifications and masses, which are common abnormalities.



**FIGURE 1.** a. Cancerous; and b. Non-cancerous breast image [3].

Microcalcifications and breast masses occur in the connective tissues and epithelia of the breast region [4]. Breast tumors emerge in the breast and differ in size and shape. These are classified as benign or malignant, depending on their severity. Benign breast lumps are non-aggressive and non-cancerous, but they expand and impinge on adjacent organs, causing additional complications [5]. Malignant breast tumors are aggressive and cancerous. They must be treated as soon as possible to avoid mortality. Benign masses are oval or circular with confined and smooth borders, whereas malignant tumors are uneven in shape. Malignant breast masses are defined as fuzzy, rough, or ambiguous lumps. Furthermore, the cancerous tumor appears whiter than any surrounding tissue. The challenges and benefits of previous breast tumor classification and detection have led to the development of an automatic technique for assisting professional radiologists in ensuring greater interpretation and accuracy.

A diagnostic mammographic image is typically pre-processed to remove the pectoral muscle with a mammogram encircling for the detection process. By removing the pectoral muscle and background areas from a mammographic image, accurate breast profile segmentation on the surface can be determined [6]. Cancer tissues with larger pixel intensities were detected more easily than those in the breast area. The intensities of opaque breasts in normal tissues are similar to those in cancer areas; hence, tumor areas are productively generated. The manual techniques implemented by radiologists fail because of the similar appearance of microcalcifications and breast masses. Finding the tumor mass by segmenting the region of interest is a challenging task in research [7]. As a result, early detection technologies combined with automated systems must aid radiologists in accurately diagnosing breast tumors.

Screening models are utilized for screening breast cancer, including clinical and self-breast checks, magnetic resource imaging (MRI), mammography, and ultrasound. Mammography is an efficient and reliable radiographic procedure for detecting breast masses [8]. During screening, a 3D model of the breast is generated from various angles. High-quality and high-resolution images are utilized in subsequent image processing techniques, including feature extraction and segmentation. Thus, prior identification of breast cancer aid in reducing the death rate was considered in this research [9]. The proposed research uses a hybrid K-means and GMM machine learning model to increase the

classification accuracy, reduce the error rate, and achieve a high signal-to-noise ratio.

The structure of this study is organized into different sections. The second section involves related works based on breast cancer classification and detection. The third Section presents the materials and methods used in the proposed work. The fourth Section discussed in detail about the experimental results in detail. The final section concludes with the novelty of this research.

## II. RELATED WORKS

The existing technique in the literature presents a computer-aided detection (CAD) method that depends on classification and feature extraction using machine learning (ML) models, which aid radiologists in identifying breast tumor lesions in X-rays. The initial process contains a pre-determined deep convolutional neural network (DCNN), and deep features are extracted in the second stage [10]. These are further fed with a support vector machine (SVM) classifier and various kernel functions. The third process presents deep feature fusion, which increases the accuracy of the SVM classifier compared to other methods.

Various methods have been used to identify various computer-aided detection approaches for breast cancer using ML techniques [11]. The inputs of these approaches are grouped into histopathological images, which have a variety of visual patterns and seem to be complicated in recognizing quality features to assist in the recognition of cancer. The author investigated various pre-trained CNNs to extract attributes from the histopathology images. These images were taken from the BreakHis dataset [12], which is publicly available.

Several approaches emphasize feature extraction, histopathological imaging, and segmentation. Pre-processing and adaptive learning based on the Gaussian aggregate model and interconnected element survey-based interest localization around the formed extraction are all components of this method. This approach operates in correlation with SVM to detect breast cancer [13].

Full-field digital mammography (FFDM) is broadly used to screen for breast cancer [14]. Contrast-enhanced digital mammography is an expanding technology in the current field comprising low-energy images related to FFDM and recombines images supporting cancer neo-angiogenesis, which are the same as breast MRI.

The advanced level of artificial intelligence (AI) technique and the natural image classification method for breast figure categorization tasks were investigated. The author has explained the performance of the neural network (NN), support vector machine (SVM), Bayesian methods, and random forest (RF) algorithms for breast image classification [15].

Advanced soft computing technology is used to pre-process the images and achieve the best classification process. Using a hybrid combination of photoacoustic images and machine learning to compare the region of the curve, the

specificity and sensitivity of SVM has the potential to have a significant impact on diagnostics [16].

A novel classification technique depends on the fuzzy Gaussian mixture model (FGMM) by merging the fuzzy logic system and Gaussian mixture model power for the CAD method. This approach is used to distinguish between normal and malignant mammography images [17]. The confusion matrix was applied to generate the FGMM performance metrics, which improved the FGMM diagnostic accuracy and reliability in breast cancer diagnosis.

Breast cancer can be detected earlier using mammography. This model is based on a technique for mammography segmentation that is given with increased thresholding [18]. Furthermore, the final segmented image from the original image can easily identify breast cancer. In general, amplified segmentation is employed in all biomedical images for better detection, feature extraction, and visualization, which improves the accuracy of diagnosis.

Fuzzy multi-layer support vector machine (FMSVM) classification was used to estimate the extracted features, and their effects were determined [19]. This method is based on a combined image set taken from the publicly available mini MIAS databases [20]. This shows the efficacy with which benign, normal, and malignant tumors can be detected. It is also used to detect the tumor area and determine the location of the tumor is mainly concentrated [21]. It focuses on identifying the best algorithms for determining the tumors that exist in the breast. The most effective strategy for tumor diagnosis is a hybrid combination of K-means, dilatation, and canny edge detection techniques.

An automated breast segmentation process is employed to find the hottest region in thermograms by employing a morphological watershed driver to assist the experts in discovering the tumor in an effective method of infrared thermography [22]. An operation for thermogram assessment is the time required to achieve the proposed thermal stabilization. Image analysis for an automated system has low breast cancer grades in digitized histopathology, and intermediates have been examined [23]. Object-level, semantic-level, pixel-level features, hematoxylin, and eosin-stained breast biopsy tissue from 106 patients were identified among the multiple levels of feature sets. In this study, a hybrid active segmentation method was used to classify nuclei from images. A cascaded approach was used to construct multiple SVM classifiers for abnormal mammogram classes [30].

A segmentation model based on various machine-learning approaches is presented [26]. This model was trained effectively using normal back propagation to improve the neural network convergence rate and segmentation. The typical technique for the segmentation process in breast cancer is discussed using an advanced soft computing paradigm [27]. Pixel-to-pixel-level classification and segmentation are effectively used to detect all mammograms. These models are effectively trained using an advanced machine-learning approach with a better accuracy rate [28]. This research is further enhanced by the gaps in the existing

soft computing strategy, which comprises the numerous tools and datasets employed in this work [29]. Breast cancer can be diagnosed at an earlier stage based on histological images. Hyper-parameter tuning was used to improve the efficiency of the trained model [31].

A residual neural network model for breast cancer segmentation is performed by fine-tuning the magnification factors. Using this process, the classification accuracy was calculated [32], [33]. Diagnostic tools are used to detect abnormalities in the breast using breast ultrasound (BUS) imaging. Three classifiers are employed to increase the classification accuracy: K-nearest neighbors (KNN), random forest, and decision tree [34], [35]. The subjective approach of classification is to use SVM and decision tree to categorize malignant and non-malignant categories [36], [37].

### III. MATERIALS AND METHODS

This study presents a K-means segmentation model using a hybrid combination approach to detect cancerous and non-cancerous breasts. For image pre-processing, an adaptive median filter was applied for K-means classification and the Gaussian mixture model (GMM). Cancer is the uncontrolled accumulation of cell groups in a specific body location and the second most common cause of death in women worldwide. It is possible to treat the condition when it is properly recognized in its early stages. Several studies have been performed to detect cancers. However, no accurate techniques have been developed to date. Hence, a novel approach was used to accurately identify tumor regions. The proposed model was utilized to visually detect tumors and determine the location of the tumor. This work mainly focuses on the detection of tumors situated in the breast and fragments benign and malignant images using K-means and GMM algorithms.

Digital mammographic images, such as normal, benign, and malignant, were obtained from the source [20]. A pre-processing technique improves the image quality for further processing by reducing or removing surplus or unrelated elements in the mammography image background.

#### A. DATASET AND DATA PREPARATION

The Mammographic Image Analysis Society (MIAS) is a consortium of UK research organizations authorized to better understand mammograms that have a digital mammography database [20]. It consists of normal and abnormal breast images of the patients. The database contains 322 open-access digitized films and is accessible on a 2.3 GB 8 mm (ExaByte) tape. The radiologist's "truth"-markings on the areas of any anomalies may be included. The database was padded/clipped and trimmed to a 200 micron pixel edge, resulting in an image size of  $1024 \times 1024$  pixels. The dataset is publicly accessible, and mammography images are acquired from the link [20]. Preprocessing is the main problem in low-level image processing. Pre-processing enhances the intensity between the background and objects, resulting in more accurate breast tissue structure projections. Screen-

film mammography (SFM) is not accurately positioned in the scanner during the digitization process. The breast area boundary is removed from background objects, such as artifacts, scanning labels, and breast position. The image was smoothed and segmented by eliminating the uneven background of the breast tissue. Therefore, accurate extraction of the breast region was achieved by deleting the boundary and background. Hence, a pre-processing technique is essential to improve the quality. It also prepares a mammogram for the forthcoming processes, namely segmentation and feature extraction. Some components, such as high frequency and noise, were removed with the assistance of an adaptive median filter.

The adaptive median filter operates in a rectangular  $xy$  space. It varies the  $R_{xy}$  size in the filtering operation based on the conditions mentioned below. The median in the 3-by-3 neighborhood near the corresponding pixel in the collected images was used to create each output pixel. The image edges, on the other hand, are replaced with zeros. The filter output holds only one value that replaces the present pixel value at  $(x, y)$ , where the point at which  $R$  is centered at time. The notation used is:

- $S_{min}$  = minimal pixel value of  $R_{xy}$
- $S_{max}$  = maximal pixel value of  $R_{xy}$
- $S_{med}$  = median pixel value of  $R_{xy}$
- $S_{xy}$  = value of pixel at coordinates  $(x, y)$
- $R_{max}$  = maximal allowed  $R_{xy}$  size

Thus, adaptive median filtering is used to smoothen the non-repulsive noise arising from 2D signals without blurring borders and conserved images. The pre-processing model is used for orientation, segmentation, label, enhancement, artifact removal, and mammography. It is used to create masks to pixels with high intensity for decreased resolution and breast segments. The median filter causes the entire image fuzzer to transform the boundaries of objects present in the image into crisp, fine, and straight lines that are isolated directly.

**Pre-processing**

Preprocessing was performed using an adaptive median filter. This is the most significant step in medical image processing for detecting breast cancer using mammography images. The pre-processing image output was utilized for noise-free image classification. Figure 2 shows the various input images, such as normal, benign, and malignant, which are considered for further processing. The boundaries between microcalcifications and breast tissue were enhanced in the initial view of the images. The outcome of an adaptive median filter shows a better restoration of grayscale images. This helps to reduce the noise level when compared to other multilevel median filter types.

**B. PROPOSED MODEL AND ALGORITHM**

The proposed model consists of an input breast database, image preprocessing, background elimination, filtering, and segmentation, as shown in Figure 3. The input dataset from the Mammographic Image Analysis Society (MIAS) is

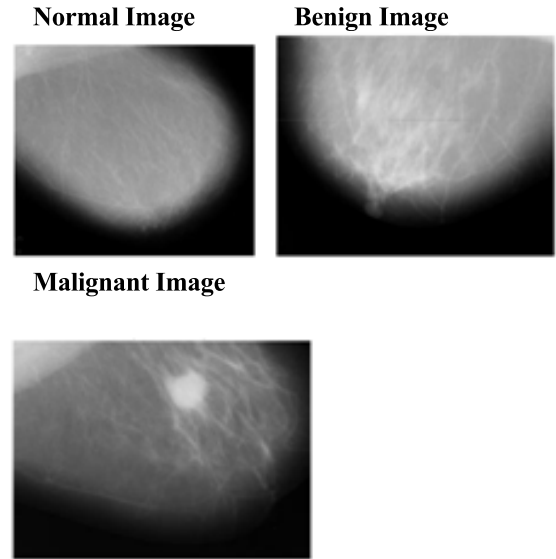


FIGURE 2. a. Normal image b. Benign image c. Malignant image [20].

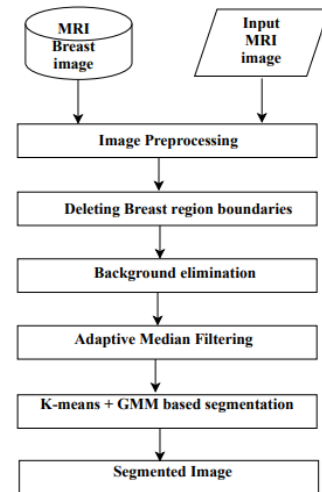


FIGURE 3. Proposed model for breast cancer segmentation.

publicly accessible, and mammography images are extracted. Low-level image processing is often used in pre-processing to increase the contrast level. This improves the intensity between the backgrounds to produce reliable breast tissue. Background elimination is the process of creating a foreground mask to separate a component from the background. This method is used to detect objects from motionless images. An adaptive median filter approach was used to remove the impulse noise and speckle from the images. In the proposed hybrid approach, the labeled features of both k-means and GMM are effectively used to partition the region or seed points into various sub-instances.

The cluster numbers and mean values were initialized using K-means. The Euclidean distance is used to determine the distance (each instant) between the center of the cluster and the case. The center of each cluster was measured using the Euclidean distance, and the instance was allocated to the cluster with the minimal distance. As a result, the

image points were labeled and clustered using the estimated distance. The cycle is terminated when each group is clustered, and each center is updated by averaging the points that belong to that cluster. When each instance permanently settles in clusters, the algorithm terminates. In other words, the instances are not transmitted from one cluster to another. GMM is a versatile segmentation approach that allows the selection of a component distribution, estimating the density for each group, and constructing soft clustered boundaries. GMM utilizes the expectation-maximization (EM) algorithm to compute the GMM parameters. The EM design is an iterative process in which the maximum likelihood is determined when the observed data are considered to be incomplete. Every frequency in the EM design contains two main processes: E-step (i.e., expectation) and M-step (maximization). In the E-step, the current estimates and observed data of the model parameters were used to evaluate the missing data. This parameter is the conditioned expectation to determine the terminology option. Under the hypothesis that such missing data are known, the M-step maximizes the probability function. The E-step was used to estimate the missing data. The design ensures that likelihood maximization occurs in each cycle, guaranteeing convergence.

GMM is a function of the likelihood to maximize the parameters, namely variance and mean. Thus, the parameters are estimated using the EM algorithm. In the initial stage, the number of means, classes, mixing coefficients, and variance were initialized. In the expectation step, compute the probabilities of the posterior with the present parameter values using (1).

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

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.

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

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

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

The log-likelihood is evaluated by (5),

$$\ln L(Y/\mu, \sigma, \pi) = \sum \ln \sum_{n=1}^N \pi_n G(x/\mu_n, \sigma_n) \quad (5)$$

According to the density calculation, the cluster k numbers in the GMM segmentation model are automatically computed using the thresholding technique for each image. The mammography images are segmented into regions of the k cluster, where every pixel belongs to a cluster after the GMM parameters are computed using the EM design. As a result, the

image is segmented into benign, normal, and malignant tissue classes using k-means and GMM. Finally, the accuracy of the segmentation method is expressed as a percentage, as in (6):

$$\begin{aligned} \text{Accuracy} &= \frac{\text{absolute TP} + \text{absolute TN}}{\text{absolute TP} + \text{absolute FP} + \text{absolute TN} + \text{absolute FN}} \\ &\times 100 \end{aligned} \quad (6)$$

where TP, TN, FN, and FP are true positive, true negative, false negative, and false positive, respectively. The above equation provides more accuracy for segmentation in the proposed method. The error rate is calculated for two n × m. Images of monochrome in Equation (7).

$$\text{Error Rate} = \frac{1}{nm} \sum_{a=0}^{n-1} \sum_{b=0}^{m-1} \|(K(a, b) - I(a, b))^2| \quad (7)$$

where K and I are two images, one of which is a noisy approximation, and the other is not. The signal-to-noise ratio (SNR) is the ratio of signal strength to noise power, which is measured and expressed in decibels (8).

$$\text{SNR}_{\text{decibel}} = 10 \log_{10} \left( \frac{R_{\text{signal}}}{R_{\text{noise}}} \right) \quad (8)$$

A signal rate greater than 1:1 (i.e., more than zero dB) indicates that the signal is greater than the noise. The steps for k-means and the GMM algorithm are as follows.

---

#### K-Means Algorithm

---

*Input: Normal, Benign, and Malignant image*

*Output: Segmented image*

*Start*

*Step 1 Set the number of clusters k to assign in the given data.*

*Step 2 Select k value randomly from the centroids of the group.*

*Step 3 Repeat*

*Step 4 Expectation: Select the each point to its closest centroid.*

*Step 5 Maximization: Estimate the new centroid of each point in the cluster.*

*Step 6 Until Centroid positions and coordinates does not change*

*End*

---



---

#### GMM Algorithm

---

*Input: Normal, Benign, and Malignant image*

*Output: Segmented image*

*Start*

*Step. 1 Consider j Gaussian classes with random μ and σ.*

*Step. 2 Calculate posterior probability of each pixel for each class*

*Step. 3 Assign pixel to class with highest probability*

*Step. 4 Update μ and σ for each class*

*Step. 5 Estimate maximum likelihood estimation*

*Step. 6 Repeat Step 2 to 6.*

*End*

---

The proposed K-means and GMM models detect breast tumors and segment the images into benign, normal, and malignant categories. Greater accuracy was obtained with a lower error rate. The pseudo-code for the proposed method is as follows:

**Proposed Algorithm**

- Input: Normal, Benign, and Malignant image*  
*Output: Segmented image*  
 Start
- Step. 1 Selecting a mammographic image from the image collection database
  - Step. 2 The pre-processing technique is applied to improve image quality
  - Step. 3 Eliminating breast region boundary and uneven background
  - Step. 4 Removal of noise and high frequency through an adaptive median filter
  - Step. 5 K-means and GMM segments data into k-clusters
  - Step. 6 Frame the expectation step using Eqn. (1)
  - Step. 7 Calculate mean, variance, and mixing coefficient in maximization step using Eqn. (2), (3), and (4)
  - Step. 8 Evaluate the log-likelihood in the GMM model using Eqn. (5)
  - Step. 9 Estimate the accuracy values using Eqn. (6)
  - Step. 10 Classification of a normal, benign, and malignant segmented image
- End

The hyper-parameters of the k-means, GMM, and hybrid methods are presented. Using this algorithm, the training process was obtained for all data in the given breast image repository. Cross-validation was used to evaluate the proposed model to determine a better breast cancer model.

**IV. EXPERIMENT, RESULTS AND DISCUSSION**

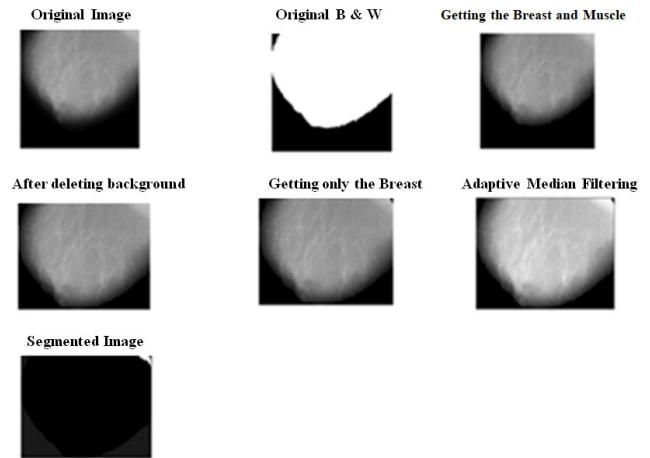
**A. MIAS DATASET**

Initially, the input data were imported from a breast data repository [20]. The original 322 images (161 pairs) at 50-micron resolution in “Portable Gray Map” (PGM) format and accompanying truth data description are included in the Mammographic Image Analysis Society (MIAS) dataset of digital mammograms (v1.21), as shown in Table 1.

**TABLE 1. Dataset descriptions.**

S.No	Statistics	Descriptions
1	Size	8 bits
2	Optical density	0 to 3.2
3	Spatial resolution	50µm pixel
4	No. of pairs	161
5	No. of images	322

A digital dataset for screening mammography (DDSM) was obtained from the University of South Florida. In image preprocessing, artifacts are one of the limitations in the given image owing to the marking of some additional lesion spots.



**FIGURE 4. Normal Image – Segmentation process flow using hybrid segmentation model.**

In addition, MIAS datasets were used to enhance the size of the data collections for further processing. Pre-processing and classification techniques were utilized to evaluate the accuracy of the proposed method (322 images, 64 benign, 51 malignant, and 207 normal breast images).

Subsequently, the images must be pre-processed to increase the difference in intensity between background objects and produce reliable breast tissue structure representations. Furthermore, an adaptive median filter was utilized to eliminate noise and high frequencies. Additionally, hybrid k-means and GMM models were applied to segment the clusters using different sets of parameters.

Input images are classified into three types, namely normal, benign, and malignant images, which also include physician marking on the place of abnormality. The database concludes with four types of abnormalities: suspicious lesions, architectural distortions, circumscribed calcifications, and masses. The proposed method was evaluated using mammography image collection, and the results are presented separately. The image set was divided into classes based on size.

**B. SEGMENTATION**

The infrared images of three different cases, namely normal, benign, and malignant, were segmented and implemented using MATLAB R2019a. When a mammographic image contains microcalcifications, the proposed method allows for binary outcomes to indicate whether the tissue is benign, normal, or malignant. This process was computed in an Intel®Core™i5–8265 U processor at 3.9 GHz using Windows®10 operating system of 64-bit with 8 GB DDR4 memory.

Figures 4, 5, and 6 depict the segmentation of normal, benign, and malignant tissues from mammogram images. A step-wise reflection of the methodology is depicted by projecting essential stages, such as removal of the pectoral muscle, filtering process, and segmentation.

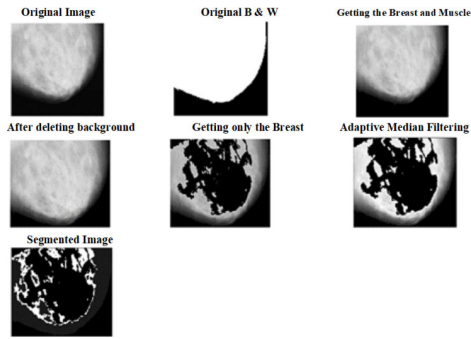


FIGURE 5. Benign Image – Segmentation process flow using hybrid segmentation model.

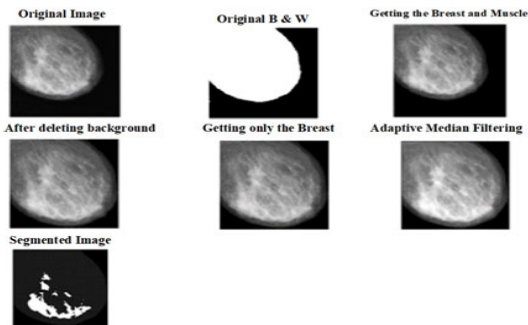


FIGURE 6. Malignant Image – Segmentation process flow using hybrid segmentation model.

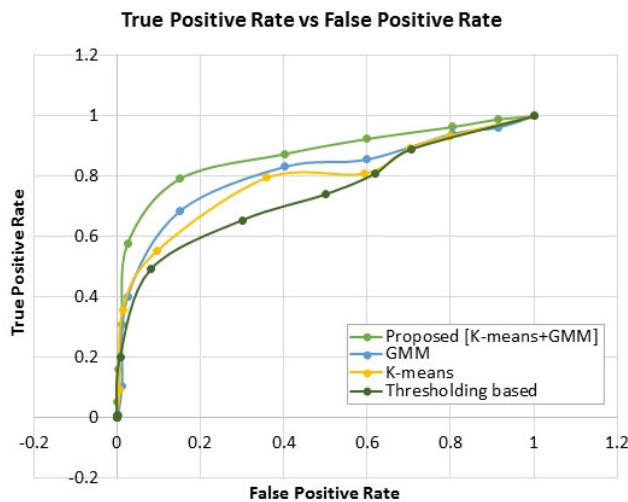


FIGURE 7. Performance comparison – TPR vs FPR.

### C. COMPARATIVE ANALYSIS

An extensive analysis of the proposed segmentation model was performed by comparing the hybrid model with three other methods: GMM, K-means, and thresholding methods. Figure 7 depicts the performance of the true-positive rate versus the false-positive rate.

Figure 8 shows that K-Means is slower than GMM with a K-Means initializer. Hybrid GMM and K-means algorithm converge after 3<sup>rd</sup> epoch. Expectation-Maximization

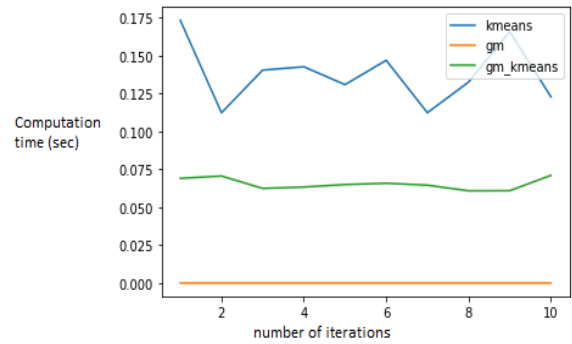


FIGURE 8. Computation time (sec) for different number of iterations.

procedure is assured to have a local maximum after 10<sup>th</sup> iteration. At this point, the overall convergence of optimized K-means and GMM is existing at 10<sup>th</sup> iteration. GMM consumes less computation time than other existing techniques. This occurs when it finds a local minimum existence that is not close to the global minimum.

When a precise value for  $k$  is specified, it can be substituted for  $k$  in the model reference, for example,  $k=10$  for 10-fold cross-validation. This method is most commonly used in applicable machine learning to determine unknown data. The region of interest (ROI) was subjected to a 10-fold cross-validation procedure. With 322 ROI images, the dataset was partitioned into 30% testing and 70 % training.

The learning rate of  $\epsilon$  is chosen via cross-validation with a value of 0.001 along with the hyper-parameter decision. It has been observed that initializing high precision to the cut-off value  $D_{max}$  and a uniform initialization of  $\tau_i$  is advantageous. The centroids were adjusted to random values. Larger values cause  $\sigma(t)$  to decline faster, which may impair convergence. Smaller values are always acceptable, but they take longer to reach convergence.

Various segmentation approaches were compared with the proposed method to validate the performance measures. K-means and GMM have 93.8% and 65% accuracy with high error rates of 29.47% and 24.35%, and low SNR, respectively. Thresholding had 86% accuracy and error rates of 32.58% and 10.17%, respectively. The accuracies of the three categories of SVM with kernel functions were 56.93%, 72.28 %, and 84.33 %, respectively. Growth region hand selection and FCM-GA selection had accuracies of 63% and 71%, respectively. The proposed hybrid model (K-Means and GMM) has a better accuracy of 95.50%, a low error rate of 18.64%, and a high SNR of 13.05. Table 2 presents a comparative analysis of classification accuracy, error rate, and SNR parameters for benign, malignant, and normal images after 10 epochs and an average execution time of 0.068 s.

The application of hybrid K-means and GMM segmentation will assist physicians in making early diagnoses by improving the qualitative identification of breast cancer in mammography images. Table 2 shows that the proposed hybrid model has a segmentation classification accuracy of

**TABLE 2. Comparative analysis of proposed model with existing techniques.**

S.No	Technique	Classification Accuracy (%)	Error Rate (%)	Signal-to-Noise Ratio (SNR)
1	<b>Proposed Hybrid Model [K Means + GMM]</b>	<b>95.50</b>	<b>18.64</b>	<b>13.05</b>
2	Gaussian Mixture Model [1]	93.80	29.47	10.23
3	K-means [1]	71.00	25.45	11.25
4	Thresholding [1]	86.00	32.58	10.17
5	SVM with kernel function (50-50) training & testing [24]	56.93	32.32	11.12
6	SVM with kernel function (60-40) training & testing [24]	72.28	19.24	12.05
7	SVM with kernel function (70-30) training & testing [24]	84.33	21.24	10.13
8	Growth region hand selection [25]	63.00	28.00	12.67
9	Growth region FCM-GA selection [25]	71.00	50.00	10.76

**TABLE 3. Multi-variant analysis of performance measures with anova test.**

	Accuracy	Error Rate	SNR
No. of samples, N	9	9	9
$\sum X$	687.84	255.84	101.67
Mean	76.43	28.43	11.29
$\sum X^2$	54126.66	8020.21	1158.18
Standard Deviation, $\sigma$	13.95	9.67	1.09

95.5 %, an error rate of 18.64, and a signal-to-noise ratio of 13.05, which is significantly more reliable than the existing techniques. Furthermore, the proposed technique minimizes the error rate.

The efficacy of the proposed method is presented in the diagnosis of breast cancer and its reliability in identifying malignant tumors from benign tumors. Using this method, medical experts can identify breast cancer faster with greater precision.

Extensive result analysis is presented with multi-variant matrices, such as accuracy, error rate, and signal-to-noise ratio. Analysis of variance (ANOVA) is a statistical approach for determining one or more variables in a set that differs significantly from one another. It checks the impact of one or more factors by comparing the means of different samples, as shown in Table 3.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \tag{9}$$

where,  $\sigma$  = standard deviation

$X_i$  = sub sets

$\bar{x}$  = arithmetic mean of data.

N = number of sample sets

$\sum (X_i - \bar{x})^2$  = Sum of all sample points.

**TABLE 4. Various statistical test.**

Test	p-value	f-ratio
ANOVA	0.045	1.0638
T-test	0.056	1.0045
Z-test	0.078	1.0003
Chi-square	0.081	0.0089

The f-ratio value was 1.0638, p-value was < 0.0001, and significant at  $p < 0.05$ .

The ANOVA test showed an improved prediction rate for the proposed breast cancer performance metrics. The hybrid model proved the improvement in the detection of malignant breast cancer. The inference of this analytical study is to improve the accuracy, lower error rate, and high SNR.

Table 4 shows that among the various statistical tests, ANOVA has better result for the proposed work.

**V. CONCLUSION AND FUTURE WORK**

In this research, two segmentation approaches, namely the K-means and Gaussian mixture model (GMM), are used to segment different categories of breast images, such as normal, benign, and malignant. It is proven that the hybrid approach has better performance measures, such as an accuracy of 95.5%, an error rate of 18.64%, and a signal-to-noise of 13.05 when compared to other existing techniques. The ANOVA test checks the impact of one or more factors by comparing the mean, variances, and standard deviations of different samples. It shows a high prediction rate for the hybrid segmentation technique used in breast cancer detection.

The hybrid GMM and K-means model is a novel method for detecting breast cancer with good accuracy. Initially, the breast image from the data repository was preprocessed.



Removal of speckle noise and special markings in medical images enhances image segmentation quality. The results show that the hybrid GMM and K-means perform better than the existing techniques. The future scope of this method shows better outcomes in terms of precision, and the segmentation models are greatly emphasized. This intelligent healthcare model will bring a revolution in the medical era by solving human problems in society, especially in detecting breast cancer in women at an early stage.

## ACKNOWLEDGMENT

The authors would like to thank all our universities for facilitating our time support in this study.

## REFERENCES

- [1] 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 classification," *Multimedia Tools Appl.*, vol. 76, no. 7, pp. 10191–10205, Apr. 2017.
- [2] M. R. Aure, V. Vitelli, S. Jernström, S. Kumar, M. Krohn, E. U. Due, T. H. Haukaas, and S. K. Leivonen, "Integrative clustering reveals a novel split in the luminal a subtype of breast cancer with impact on outcome," *Breast Cancer Res.*, vol. 19, no. 1, pp. 1–18, Dec. 2017.
- [3] S. Kaymak, A. Helwan, and D. Uzun, "Breast cancer image classification using artificial neural networks," *Proc. Comput. Sci.*, vol. 120, pp. 126–131, Jan. 2017.
- [4] S. M. Badawy, A. A. Hefnawy, H. E. Zidan, and M. T. GadAllah, "Breast cancer detection with mammogram segmentation: A qualitative study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 1–4, 2017.
- [5] Z. Wang, L. Zhang, X. Shu, Q. Lv, and Z. Yi, "An end-to-end mammogram diagnosis: A new multi-instance and multiscale method based on single-image feature," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 3, pp. 535–545, Sep. 2021.
- [6] E. H. Cain, A. Saha, M. R. Harowicz, J. R. Marks, P. K. Marcom, and M. A. Mazurowski, "Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: A study using an independent validation set," *Breast Cancer Res. Treatment*, vol. 173, no. 2, pp. 455–463, Jan. 2019.
- [7] A. H. Yurtakal, H. Erbay, and T. İkizceli, "Detection of breast cancer via deep convolution neural networks using MRI images," *Multimedia Tools Appl.*, vol. 79, pp. 15555–15573, Apr. 2019.
- [8] J. Zhang, B. Chen, M. Zhou, H. Lan, and F. Gao, "Photoacoustic image classification and segmentation of breast cancer: A feasibility study," *IEEE Access*, vol. 7, pp. 5457–5466, 2019.
- [9] R. Chen, W. Wu, H. Qi, J. Wang, and H. Wang, "A stacked autoencoder neural network algorithm for breast cancer diagnosis with magnetic detection electrical impedance tomography," *IEEE Access*, vol. 8, pp. 5428–5437, 2020.
- [10] F. Gao, T. Wu, J. Li, L. Ruan, D. Shang, and B. Patel, "SD-CNN: A shallow-deep CNN for improved breast cancer diagnosis," *Comput. Med. Imag. Graph.*, vol. 70, pp. 53–62, Dec. 2018.
- [11] M. Garduño-Ramón, S. Vega-Mancilla, L. Morales-Henández, and R. Osornio-Rios, "Supportive noninvasive tool for the diagnosis of breast cancer using a thermographic camera as sensor," *Sensors*, vol. 17, no. 3, p. 497, Mar. 2017.
- [12] *Robotic Vision and Imaging Laboratory*. Accessed: Oct. 1, 2020. [Online]. Available: <https://web.inf.ufpr.br/vri/databases/>
- [13] V. Hariraj, W. Khairunizam, V. Vikneswaran, Z. Ibrahim, A. B. Shahrman, and M. R. Zuradzman, "Fuzzy multi-layer SVM classification of breast cancer mammogram images," *Int. J. Mech. Eng. Tech.*, vol. 9, no. 8, pp. 1281–1299, 2018.
- [14] R. F. Mansour, "A robust deep neural network based breast cancer detection and classification," *Int. J. Comput. Intell. Appl.*, vol. 19, no. 1, Mar. 2020, Art. no. 2050007.
- [15] I. L. Milankovic, N. V. Mijailovic, N. D. Filipovic, and A. S. Peulic, "Acceleration of image segmentation algorithm for (Breast) mammogram images using high-performance reconfigurable dataflow computers," *Comput. Math. Methods Med.*, vol. 2017, May 2017, Art. no. 7909282.
- [16] A.-A. Nahid and Y. Kong, "Involvement of machine learning for breast cancer image classification: A survey," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–29, 2017.
- [17] I. Prabhakaran, Z. Wu, C. Lee, B. Tong, S. Steeman, G. Koo, P. J. Zhang, and M. A. Guvakova, "Gaussian mixture models for probabilistic classification of breast cancer," *Cancer Res.*, vol. 79, no. 13, pp. 3492–3502, Jul. 2019.
- [18] D. A. Ragab, O. Attallah, M. Sharkas, J. Ren, and S. Marshall, "A framework for breast cancer classification using multi-DCNNs," *Comput. Biol. Med.*, vol. 131, Apr. 2021, Art. no. 104245.
- [19] S. Saxena, S. Shukla, and M. Gyanchandani, "Pre-trained convolutional neural networks as feature extractors for diagnosis of breast cancer using histopathology," *Int. J. Imag. Syst. Technol.*, vol. 30, no. 3, pp. 577–591, Sep. 2020.
- [20] *Mammographic Image Analysis Homepage*. Accessed: Oct. 1, 2020. [Online]. Available: <https://www.mammoimage.org/databases/>
- [21] P. M. Shakeel, T. E. El Tobely, H. Al-Feel, G. Manogaran, and S. Baskar, "Neural network based brain tumor detection using wireless infrared imaging sensor," *IEEE Access*, vol. 7, pp. 5577–5588, 2019.
- [22] R. Turkki, D. Bychkov, M. Lundin, J. Isola, S. Nordling, P. E. Kovanen, C. Verrill, K. von Smitten, H. Joensuu, J. Lundin, and N. Linder, "Breast cancer outcome prediction with tumour tissue images and machine learning," *Breast Cancer Res. Treatment*, vol. 177, no. 1, pp. 41–52, Aug. 2019.
- [23] T. Wan, J. Cao, J. Chen, and Z. Qin, "Automated grading of breast cancer histopathology using cascaded ensemble with combination of multi-level image features," *Neurocomputing*, vol. 229, pp. 34–44, Mar. 2017.
- [24] T. A. Shaikh and R. Ali, "An intelligent healthcare system for optimized breast cancer diagnosis using harmony search and simulated annealing (HS-SA) algorithm," *Informat. Med. Unlocked*, vol. 21, Mar. 2020, Art. no. 100408.
- [25] Z. N. Isfahani, I. Jannat-Dastjerdi, F. Eskandari, S. J. Ghouschi, and Y. Pourasad, "Presentation of novel hybrid algorithm for detection and classification of breast cancer using growth region method and probabilistic neural network," *Comput. Intell. Neurosci.*, vol. 2021, Jun. 2021, Art. no. 5863496.
- [26] C. Zhao, R. Shuai, L. Ma, W. Liu, and M. Wu, "Segmentation of dermoscopy images based on deformable 3D convolution and ResU-Net++," *Med. Biol. Eng. Comput.*, vol. 59, no. 9, pp. 1815–1832, Sep. 2021, doi: [10.1007/s11517-021-02397-9](https://doi.org/10.1007/s11517-021-02397-9).
- [27] C. Zhao, R. Shuai, L. Ma, W. Liu, D. Hu, and M. Wu, "Dermoscopy image classification based on StyleGAN and DenseNet201," *IEEE Access*, vol. 9, pp. 8659–8679, 2021, doi: [10.1109/ACCESS.2021.3049600](https://doi.org/10.1109/ACCESS.2021.3049600).
- [28] K. B. Soulam, N. Kaabouch, M. N. Saidi, and A. Tamtaoui, "Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102481, doi: [10.1016/j.bspc.2021.102481](https://doi.org/10.1016/j.bspc.2021.102481).
- [29] Z. Rezaei, "A review on image-based approaches for breast cancer detection, segmentation, and classification," *Expert Syst. Appl.*, vol. 182, Nov. 2021, Art. no. 115204, doi: [10.1016/j.eswa.2021.115204](https://doi.org/10.1016/j.eswa.2021.115204).
- [30] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, and A. A. Basha, "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform," *Measurement*, vol. 146, pp. 800–805, Nov. 2019, doi: [10.1016/j.measurement.2019.05.083](https://doi.org/10.1016/j.measurement.2019.05.083).
- [31] K. Kousalya and T. Saranya, "Improved the detection and classification of breast cancer using hyper parameter tuning," *Mater. Today, Proc.*, pp. 1–6, 2021, doi: [10.1016/j.matpr.2021.03.707](https://doi.org/10.1016/j.matpr.2021.03.707).
- [32] V. Gupta, M. Vasudev, A. Doegar, and N. Sambyal, "Breast cancer detection from histopathology images using modified residual neural networks," *Biocybernetics Biomed. Eng.*, vol. 41, no. 4, pp. 1272–1287, Oct. 2021, doi: [10.1016/j.bbe.2021.08.011](https://doi.org/10.1016/j.bbe.2021.08.011).
- [33] S. Pavithra, R. Vanithamani, and J. Justin, "Computer aided breast cancer detection using ultrasound images," *Mater. Today, Proc.*, vol. 33, pp. 4802–4807, Oct. 2020, doi: [10.1016/j.matpr.2020.08.381](https://doi.org/10.1016/j.matpr.2020.08.381).
- [34] N. Goyal and M. C. Trivedi, "Breast cancer classification and identification using machine learning approaches," *Mater. Today, Proc.*, 2020, pp. 1–4, doi: [10.1016/j.matpr.2020.10.666](https://doi.org/10.1016/j.matpr.2020.10.666).
- [35] M. A. Al-antari, S.-M. Han, and T.-S. Kim, "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105584, doi: [10.1016/j.cmpb.2020.105584](https://doi.org/10.1016/j.cmpb.2020.105584).

- [36] V. Rajinikanth, S. Kadry, D. Taniar, R. Damasevicius, and H. T. Rauf, "Breast-cancer detection using thermal images with marine-predators-algorithm selected features," in *Proc. 7th Int. Conf. Bio Signals, Images, Instrum. (ICBSII)*, Mar. 2021, pp. 1–6, doi: [10.1109/ICBSII51839.2021.9445166](https://doi.org/10.1109/ICBSII51839.2021.9445166).
- [37] R. Irfan, A. A. Almazroi, H. T. Rauf, E. A. Nasr, and A. E. Abdelgawad, "Dilated semantic segmentation for breast ultrasonic lesion detection using parallel feature fusion," *Diagnostics*, vol. 11, no. 7, p. 1212, 2021, doi: [10.3390/diagnostics11071212](https://doi.org/10.3390/diagnostics11071212).



**P. ESTHER JEBARANI** received the bachelor's degree in computer applications from Bharathiyar University, Coimbatore, and the master's degree in computer applications from the Karunya Institute of Technology and Sciences, Coimbatore. She is currently pursuing the Ph.D. degree in medical image processing and advanced soft computing paradigm with the Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science. She is working as an Assistant Professor with the

Department of Computer Science, Kovai Kalaimagal College of Arts and Science, Coimbatore. She has published various research articles indexed in Scopus. Her research interests include digital image processing, machine learning, and networking.



**N. UMADEVI** received the Ph.D. degree in computer science from Avinashilingam University for Women, Coimbatore. She is currently working as an Associate Professor and the Head with the Department of Computer Science and Information Technology, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore. She had more than 20 years of teaching experience and three years of industrial experience. She has published around 20 research

papers in both national and international journals and conferences. Around 25 M.Phil. scholars were awarded under her guidance. At present, she is guiding six Ph.D. scholars. Her research interests include image processing and data mining.



**HIEN DANG** (Member, IEEE) received the Ph.D. degree in computer science, in 2010. She is currently a Research Scholar at the University of Massachusetts Boston, USA. At the same time, she is also working at the Faculty of Engineering and Computer Science, Thuyloi University, Vietnam. She is the author of three books and more than 30 articles. In addition, she is a Team Leader or a member of many national, ministerial, and corporate projects to solve real-world problems. Her research interests include artificial intelligence, artificial neural networks, deep learning, big data analytics, and some healthcare problems.



**MARC POMPLUN** is currently a Professor of computer science at the University of Massachusetts Boston and the Director of the Visual Attention Laboratory. His work focuses on analyzing, modeling, and simulating aspects of human vision.

...