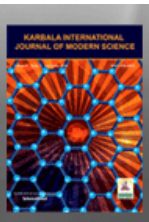
Karbala International Journal of Modern Science

Manuscript 3385

Breast Cancer Area Identification in Mammograms Using Expectation Maximization Gaussian Mixture Model

Rizki Khoirun Nisa

Dian Kurniasari

Favorisen R. Lumbanraja

Warsono Warsono

Breast Cancer Area Identification in Mammograms Using Expectation Maximization Gaussian Mixture Model

Abstract

Breast cancer accounts for 25% of all cancer diagnoses and 16% of cancer-related deaths among women globally, with high mortality rates due to late diagnosis. Early detection relies on imaging techniques such as mammography, histopathology, and breast ultrasound, with mammography being the gold standard due to its proven to detect breast cancer, thus it is effective for breast cancer treatment. However, mammogram images often produce noise and artefacts, complicating early-stage cancer detection and emphasizing the need for advanced image processing. Clustering algorithms such as K-means and Expectation Maximization - Gaussian Mixture Model (EM-GMM) have shown potential in image segmentation. This study investigates the application of EM-GMM for mammogram image segmentation, a relatively underexplored area compared to K-means, focusing on identifying potentially cancerous regions. EM-GMM, which models data based on Gaussian distribution through iterative EM algorithm application, was used on grayscale images of malignant tissue. The optimal cluster number for detecting cancerous areas was determined to be nine using the Bayesian Information Criterion (BIC). This was tested and compared with segmentations using seven and eight clusters. Results demonstrate that the segmentation with nine-cluster achieves the highest accuracy for segmenting and identifying cancerous regions. These findings support early detection and thorough breast health assessment, potentially enhancing diagnostic precision and lowering breast cancer mortality and morbidity.

Keywords

Breast cancer; Early Detection; Mammography segmentation; BIC; EM-GMM

RESEARCH PAPER

Breast Cancer Area Identification in Mammograms Using Expectation Maximization Gaussian Mixture Model

Rizki K. Nisa [a], Dian Kurniasari [a,*], Favorisen R. Lumbanraja [b], Warsono Warsono [a]

[a] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Bandar Lampung, Indonesia
[b] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Bandar Lampung, Indonesia

Abstract

Breast cancer accounts for 25 % of all cancer diagnoses and 16 % of cancer-related deaths among women globally, with high mortality rates due to late diagnosis. Early detection relies on imaging techniques such as mammography, histopathology, and breast ultrasound, with mammography being the gold standard due to its proven to detect breast cancer, thus it is effective for breast cancer treatment. However, mammogram images often produce noise and artefacts, complicating early-stage cancer detection and emphasizing the need for advanced image processing. Clustering algorithms such as K-means and Expectation Maximization - Gaussian Mixture Model (EM-GMM) have shown potential in image segmentation. This study investigates the application of EM-GMM for mammogram image segmentation, a relatively underexplored area compared to K-means, focusing on identifying potentially cancerous regions. EM-GMM, which models data based on Gaussian distribution through iterative EM algorithm application, was used on grayscale images of malignant tissue. The optimal cluster number for detecting cancerous areas was determined to be nine using the Bayesian Information Criterion (BIC). This was tested and compared with segmentations using seven and eight clusters. Results demonstrate that the segmentation with nine-cluster achieves the highest accuracy for segmenting and identifying cancerous regions. These findings support early detection and thorough breast health assessment, potentially enhancing diagnostic precision and lowering breast cancer mortality and morbidity.

Keywords: Breast cancer, Early detection, Mammography segmentation, BIC, EM-GMM

1. Introduction

Breast cancer remains the leading cause of cancer-related mortality among women globally, accounting for one in four cancer diagnoses and one in six cancer-related deaths. The areas with the highest elevated incidence rates are France, Australia/New Zealand, North America, and Northern Europe. These regions experience rates four times greater than those in Central-Southern Asia and Central Africa. Since 2022, breast cancer became the second most prevalent cancer worldwide, with approximately 2.3 million new cases, constituting 11.6 % of all cancer diagnoses. The elevated mortality rate is attributable primarily to delays in early detection.

Early detection is crucial for preventing mortality because of breast cancer. Research indicates that prompt identification and accurate diagnosis can significantly enhance patient outcomes, potentially reducing mortality rates by up to 25 % [1–8]. As a result, various imaging modalities have been developed to detect breast cancer at its early stages. These include mammography, histopathology, Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), thermography, Computed Tomography (CT), breast ultrasound (USG), and digital breast tomosynthesis [9]. The literature comprehensively reviews these techniques in detail [10–12].

Mammography is widely recognized as a primary technique for early detection due to its proven

Received 1 August 2024; revised 21 November 2024; accepted 23 November 2024.
Available online 19 December 2024

* Corresponding author.
E-mail address: dian.kurniasari@fmipa.unila.ac.id (D. Kurniasari).

https://doi.org/10.33640/2405-609X.3385

effectiveness of detecting breast cancer. Thus reducing mortality and morbidity rates of breast cancer patients [13–20]. This method utilizes X-rays to produce grayscale images called mammograms. However, these images may also produce noise and background artefacts, which can complicate the detection and interpretation of early-stage cancers [21]. Thus, advanced image processing techniques are required to enhance diagnostic accuracy.

In image processing, a pivotal technique is the segmentation of mammography images. This method is important in accurately identifying cancerous regions within mammograms, as a result, improving diagnostic precision [22]. Segmentation involves partitioning image into distinct segments or objects based on the same characteristics, such as colour, intensity levels, or texture. The primary aim of this process is to delineate a Region of Interest (ROI), thereby facilitating a more streamlined identification of breast abnormalities by radiologists [6].

Several mathematical and statistical techniques are utilized in image processing, with clustering-based methods excelling notably in image segmentation. These methods are applicable to both annotated and unannotated datasets [23]. Two of the most prevalent algorithms in this domain are K-means and Expectation Maximization with Gaussian Mixture Models (EM-GMM). Both approaches involve iterative refinement to achieve optimal clustering outcomes. Specifically, K-means uses the Euclidean distance metric for data point comparison, while EM-GMM implements a statistical methodology [24].

Previous research has thoroughly studied the application of various methods for processing breast cancer images, with a significant emphasis on the K-means algorithm. For instance, A. Arjmand, A. Farzamnia, R. Afrouzian, and S. Meshgini [25] proposed a clustering-based technique for automatically segmenting tumours in MRI scans. Their approach combined the K-means clustering algorithm with Cuckoo Search Optimization (CSO) for improved centroid initialization. Evaluation of the RIDER breast dataset indicated that their algorithm outperformed the conventional Fuzzy C-Means (FCM) algorithm. In another study, J. Dabass, M. Hanmandlu, and R. Vig [26] focused on pre-processing and segmenting breast density. They utilized binarization and modified region-growing techniques to remove labels, background noise, and pectoral muscles, followed by K-means clustering to segment digital mammograms into different density regions. Validation using the Mini-MIAS database and radiologist assessments demonstrated that their method achieved competitive results compared to current state-of-the-art techniques.

J. S. Isaac, K. Priya, M. J. Kumar, S. Kottu, V. Senthilkumar, and V. S. Ramakrishna [27] researched breast cancer classification from ultrasound images. Their study utilized both raw ultrasound scans and cancer-masked images obtained from Kaggle. Before input into the segmentation model, the images underwent pre-processing, including resizing and filtering. The K-means algorithm was also used to identify cancerous regions within the ultrasound images. Further investigations into applying the K-means algorithm in similar contexts can be found in Refs. [28–34].

In recent years, the application of EM-GMM in image segmentation, particularly in mammography, has been relatively overlooked compared to the more widely studied K-means algorithm. Over the past decade has seen a scarcity of research in this area. Notable contributions include those by A. S. Begum, K. Bhuvaneshwari, K. J. Sri, M. Divya, and R. M. Prakash [35], which their research aim is to evaluate the performance of three segmentation techniques—K-means, FCM, and EM-GMM—for infrared breast image segmentation. A. H. Yurttakal, G. Çinarer, H. Erbay, S. Karaçavuş, and T. İkizceli [36] conducted a comparative study of various segmentation and classification methods for breast cancer MRI. Furthermore, I. Khoulqi, N. Idrissi, and M. Sarfraz [37] concentrated on pectoral muscle segmentation in mammography using EM-GMM. This study aim to address this research gap by further investigating the potential of EM-GMM for mammogram segmentation, specifically on accurately identifying regions that are indicative of suspected malignant tumours.

The comparison between the previously existing research methods and the recently proposed method is presented in summary form in Table 1.

2. Material and methods

The research methodology is structured into several distinct phases, commencing with image selection and data input. This initial stage is followed by image pre-processing, which consists of resizing and histogram equalization to enhance image quality. Next, the EM-GMM step employs the EM-GMM to identify the optimal number of clusters required for effective segmentation. This step leads to the ROI process.

In the ROI process, the segmented data is undergoes further refinement through thresholding and morphological operations to achieve optimal results. Each phase is crucial for ensuring accurate

Table 1. A comparative analysis and overview of relevant research.

Researchers	Imaging Modalities	Methods and Objectives of Segmentation
A. Arjmand, A. Farzamnia, R. Afrouzian, and S. Meshgini [25]	MRI using the RIDER breast cancer dataset	K-Means + CSO to determine tumour masses in breast MRI images
J. Dabass, M. Hanmandlu, R. Vig [26]	Mammogram in Mini-Mias	Application of K-means clustering for categorizing digital mammograms into distinct density regions
J. S. Isaac, K. Priya, M. J. Kumar, S. Kottu, V. Senthilkumar, and V. S. Ramakrishna [27]	Ultrasonography	Concept of Machine Learning for classification, while K-Means is utilized as a component of the segmentation procedure to identify the location of cancer.
F. G. Y. Ciklacandir, A. Ertaylan, U. Binzat, A. Kut [28]	Ultrasonography	K-Means algorithm is implemented for lesion detection—defined as any form of damage or abnormal alteration in body tissue, which may be benign or malignant—in breast imaging by applying three distinct filters: Median, Laplace, and Sobel.
P. P. Golagani, S. K. Beebi [29]	Signal transduction pathways like PKB, MAPK, MTOR, FasL, Notch, SHH, Tnf and Wnt	K-means clustering algorithm is utilized to generate clusters from the pre-processed data. This study examines variations in metabolic pathways within breast cancer cells and identifies possible alterations to these pathways.
H. Lin, and Z. Ji [30]	Open-source Wisconsin breast cancer data set in the machine learning library of UCI	A hybrid K-Means and Self-Organizing Map model has been employed to classify breast cancer datasets into two distinct categories: benign and malignant.
D. N. Ouedraogo [31]	Breast cancer data from the UC Irvine Machine Learning Repository	Interpretable machine learning techniques such as LIME, ELI5, and SHAP, alongside machine learning algorithms, are utilized for tumour-type prediction. The K-means clustering algorithm is applied to the generated outputs.
S. H. Abdulla, A. M. Sagheer, H. Veisi [32]	Breast imaging subset of DDSM (CBIS-DDSM)	Classification of microcalcifications (MC) is proposed through a two-tier segmentation process. Initially, the breast area is extracted from the image using k-means clustering. Subsequently, an optimized region growing (ORG) approach is employed, wherein multi-seed points and thresholds are generated optimally based on the pixel intensity values in the image.
K. Wisaeng [33]	Mammogram in Mini-Mias, DDSM, and BCDR	K-Means++CSO for enhanced breast cancer detection through lesion tracking on mammographic surfaces
L. Panigrahi, R. R. Panigrahi [34]	Ultrasonography	Enhance the traditional K-means algorithm for breast cancer image segmentation by integrating Ant Colony Optimization (ACO) for initializing cluster centres and substituting Euclidean Distance (ED) with Manhattan Distance (MD). This approach seeks to maximize area preservation in the segmented images.
A. S. Begum, K. Bhuvaneshwari, K. J. Sri, M. Divya, and R. M. Prakash [35]	Infrared breast images	Three segmentation techniques, K-Means, Fuzzy C- Means (FCM) and EM-GMM, are employed to segment and compare the IR breast images. The method is applied to classify the malignant and benign cancer tissues.
A. H. Yurttakal, G. Çinarer, H. Erbay, S. Karaçavuş, and T. İkizceli [36]	**MRI Image**	The EM algorithm is exclusively employed for the generation of prototypes from previously analysed area segments.
I. Khoulqi, N. Idrissi, and M. Sarfraz [37]	Mammogram in Mini-Mias	EM-GMM segmentation for Computer-Aided Diagnosis (CAD) System
Proposed Method	Mammogram in Mini-Mias	Clustering-based methods, specifically EM-GMM, are employed to determine the optimal clusters for identifying specific tissues or areas associated with cancerous or malignant tumours.

segmentation of breast cancer images, thereby establishing a systematic approach for the identification and analysis of malignant regions. The research methodology is visually represented in Fig. 1.

2.1. Dataset

The dataset used in this study comprises grayscale mammogram images focused on breast cancer. These images were obtained from the breast cancer

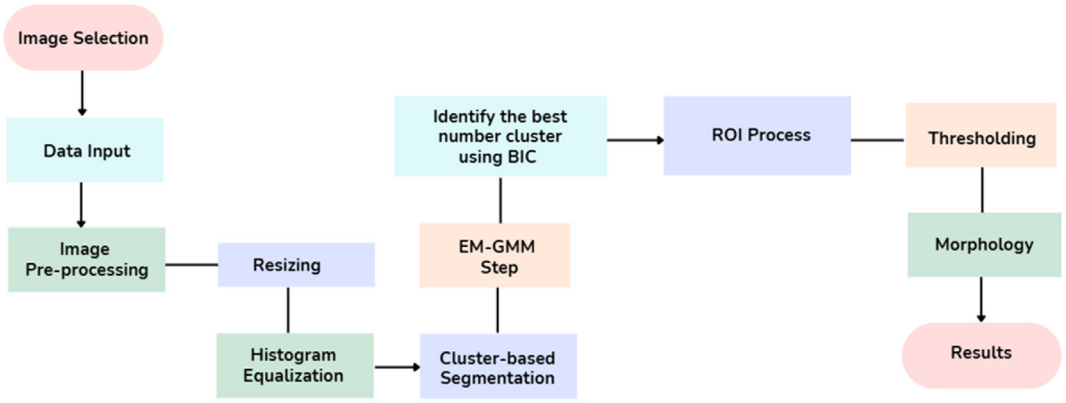

Fig. 1. Flowchart of research methodology.

mammogram database available through the PEIPA website (http://peipa.essex.ac.uk/info/mias.html) and are supported by the European project "Performance Characterization in Computer Vision" directed by Patrick Courtney at Visual Automation in Manchester, UK. The dataset consists of 322 images, subdivided into 207 normal cases and 63 benign and 52 malignant tumours, all in Portable Gray Map (PGM) format. The fundamental difference between benign and malignant tumours resides in the behaviour of their cellular compositions. Benign tumours are consist of cells that do not threaten adjacent tissue invasion, while malignant tumours contain cancerous cells that exhibit uncontrolled proliferation and the capacity to invade surrounding tissues. Therefore, this study focused solely on the analysis of 52 images of malignant tumours during the segmentation process applying the EM-GMM method, as images identified as cancerous are most relevant to the objectives of this research. This research aims to identify particular tissues or regions that have experienced malignant transformation or cancerous alterations.

2.2. Image pre-processing

In mammogram image analysis, the term "pre-processing" refers to a crucial initial step aimed at improving the quality and reliability of the results. The pre-processing steps in this study include image resizing and histogram equalization. Image resizing involves dynamically adjusting the dimensions of images to enhance their visual clarity under various conditions [38–40]. The transformation generated by the resizing process produces a new image that is scaled according to a pre-defined scale factor, commonly described by Equation (1).

$$x' = x \,.\, S_x, y' = y \,.\, S_y \tag{1}$$

where x' and y' are the new (resized) coordinates, x and y are the original coordinates, S_x and S_y are the scaling factors in the x- and y-directions, respectively.

After the resizing process, the image is transformed using histogram equalization, a well-established technique used to enhance overall image contrast by redistributing pixel intensities [41]. This step is crucial because grayscale images often exhibit low brightness levels, which typically produce low contrast or excessively bright images that display uneven histogram distributions. Therefore, to achieve a histogram distribution that aligns with our objectives, it is necessary to modify the distribution of intensity values within the image.

Performing histogram equalization requires the use of a cumulative distribution function (CDF), which represents the cumulative sum of the histogram values. The cumulative distribution function is provided in Equation (2).

$$f(k) = \frac{(N-1)}{M} . \sum_{k=0}^{n} h(k) \;\; ; n = 1, 2, 3, \ldots \tag{2}$$

where M denotes the pixels, N signifies the grayscale values, and k indicates the histogram corresponding to a specific grey value.

2.3. Cluster-based segmentation

Two most known methods in cluster-based image segmentation are spectral clustering and soft clustering. Spectral clustering is highly regarded for its effectiveness in this area. However, numerous studies [42–46] have highlighted its vulnerability to noise and outliers. This study utilises a soft

clustering technique, EM-GMM, as an alternative. This method allows each pixel in the image to be associated with several clusters, each with different membership levels, which can potentially improve segmentation precision and overall effectiveness.

2.3.1. Gaussian mixture model

The GMM is a clustering algorithm grounded in mixture models, initially introduced by Wolfe [47] and further studied by G. J. McLachlan and K. E. Basford [48]. GMM represents each cluster as a Gaussian distribution, with the overall model being a finite mixture of these Gaussian components. The defining parameters for each cluster include the mean, covariance, and weight, which are estimated using the Expectation-Maximization (EM) algorithm [49,50]. The EM algorithm and its variants are the predominant methods for estimating the Maximum Likelihood Estimate (MLE) of GMM parameters [51].

2.3.2. Expectation maximization

The EM algorithm is a robust iterative technique employed to estimate model parameters by maximizing the likelihood function, commonly known as the Q-function. It is widely used to compute parameters' MLE in probabilistic models with latent variables, such as GMM [52].

The EM algorithm consists of two main stages: the Expectation (E) step, where the expected value of the complete dataset—including any missing data—is computed, and the Maximization (M) step, where the Q-function is optimized concerning the unknown parameters. This iterative procedure persists until the convergence criteria are met, thereby ensuring precise estimation of parameters [53].

In the EM algorithm for mixed models, the complete data is represented as $x_i' = (y_i', z_i')$, where $z_i' = (z_{i1}, z_{i2}, \ldots z_{ik})$ denotes latent variables. Here, z_{ik} indicates whether x_i' associated with cluster G_k or not. The EM algorithm is formally defined by Equation (3):

$$z_{ik} = \begin{cases} 1 \ \ldots\ldots\ldots\ldots\ldots\ldots x_i \in G_k \\ 0, \ldots\ldots\ldots\ldots\ldots\ldots otherwise \end{cases} \tag{3}$$

2.3.3. Bayesian information criterion

During this stage, the clustering model is assessed using the Bayesian Information Criterion (BIC) to identify the most suitable number of clusters by evaluating the parameters of each model. The BIC balances the trade-off between model complexity and its ability to categorize the data effectively. Furthermore, this stage guarantees that the chosen optimal model aligns with the characteristics of the dataset, thereby offering a cohesive representation of the overall model [54]. The model with the highest BIC value will be chosen as optimal.

2.4. Region of interest

Measuring ROI is a fundamental and widely used process in medical image interpretation across various imaging modalities. An ROI refers to a precisely defined region within an image that is carefully analyzed by medical experts [55]. In this study, ROI selection will be performed after image pre-processing, utilizing thresholding techniques and morphological operations.

2.4.1. Thresholding

As defined by R. C. Gonzalez and R. E. Woods [56], thresholding is a process that converts an image into a binary format, using two grey levels—black and white—based on pixel values relative to a specified threshold (T). Pixels with values greater than T are rendered white, while those less than or equal to T are rendered black. This technique is effective for separating objects from the background in grayscale images, with white representing the objects and black representing the background.

Various thresholding techniques have been developed [57–60], including global methods based on histograms, local feature-based approaches, and adaptive thresholding, which we applied in our study. This adaptive method allows for tailored threshold adjustments in different image segments, enhancing object segmentation and effectively separating small regions from the background. In general, the thresholding process can be concisely represented by Equation (4):

$$g(x,y) = \begin{matrix} \mathbf{1} & \boldsymbol{if}\ \boldsymbol{f}(x,y) \geq \boldsymbol{T} \\ \mathbf{0} & \boldsymbol{if}\ \boldsymbol{f}(x,y) \geq \boldsymbol{T} \end{matrix} \tag{4}$$

2.4.2. Morphology

Morphology requires a set of nonlinear operations designed to alter the shape or structure of image features. These operations are predicated on pixel arrangement rather than their numerical values, rendering them particularly effective for processing binary and grayscale images [61].

Fundamental to mathematical morphology are two principal operators: the opening operation, which involves erosion followed by dilation. Erosion reduces the size of objects in an image, while dilation enlarges them. This operator is formally defined by Equation (5):

$$\boldsymbol{A o B} = = (\boldsymbol{A \Theta B}) \oplus \boldsymbol{B} \tag{5}$$

3. Experiment and results

3.1. Input data and resizing

The process begins with digital image processing of grayscale breast cancer mammograms. Initially stored in *PGM format, these images are converted to *JPG format before being imported into R Studio software. The images are resized to a standard 256 × 256 pixels before processing to ensure uniformity. Fig. 2 displays twenty-six of these resized samples.

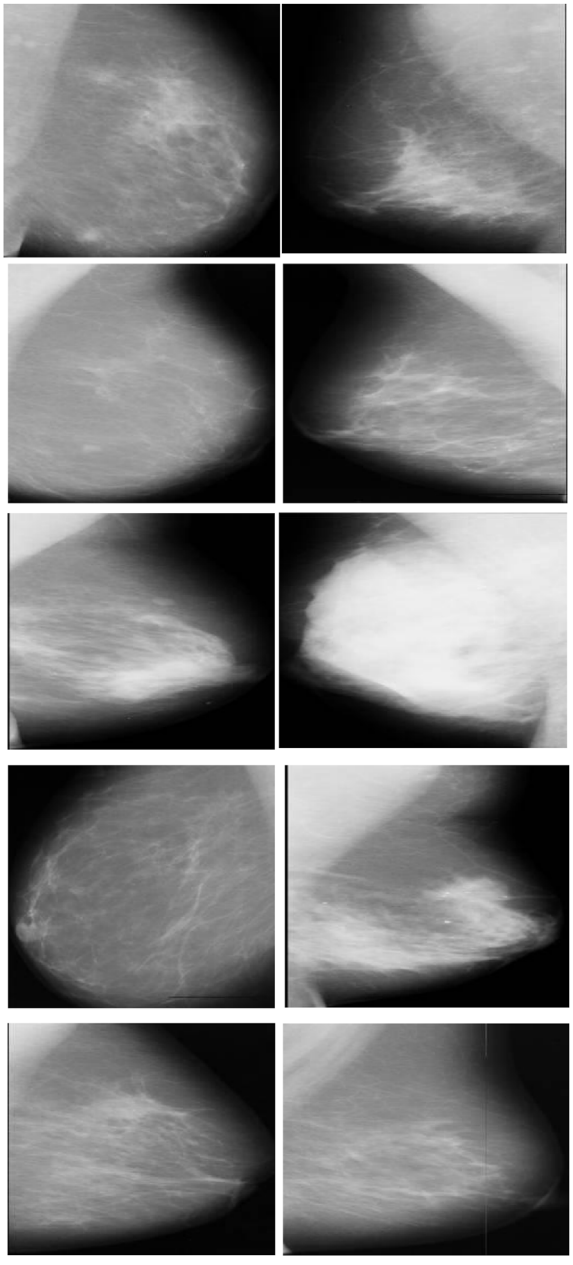

Fig. 2. Resized mammogram results.

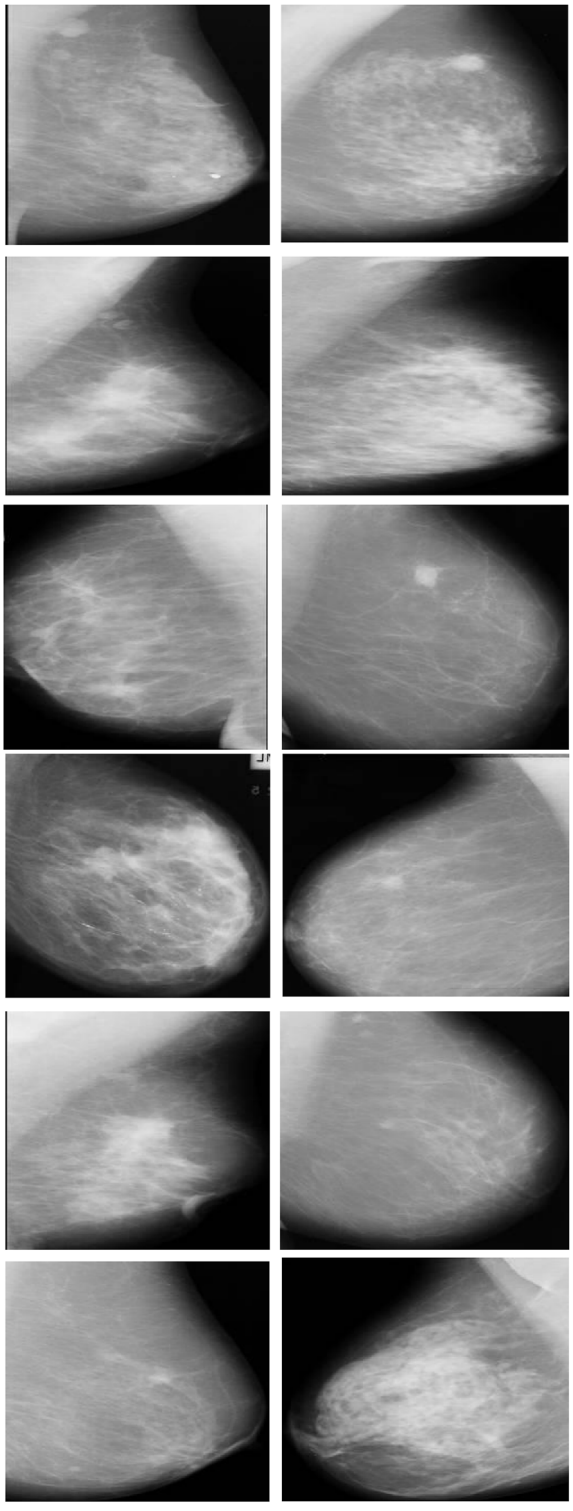

Fig. 2. (Continued).

The original and resized images show minimal variance during the resizing process, as only the pixel arrangement is altered. This step is crucial for efficient image processing within R Studio. If the image's pixel dimensions are too large, it may cause

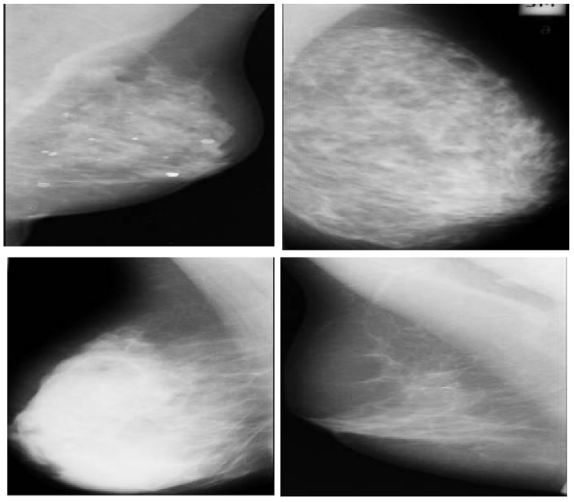

Fig. 2. (Continued)

to processing difficulties that could compromise the program's functionality.

3.2. Histogram equalization

The next step in image processing is applying histogram equalization to represent the image. Images with dullness, low contrast, or excessive brightness often have histograms with uneven distributions, as shown in Fig. 3.

Fig. 3 displays a histogram of the original image, indicating a non-uniform pixel intensities distribution. Histogram equalization is employed to achieve a more uniform distribution of grey levels, ensuring that each grey level represents approximately the same number of pixels. This technique is used to enhance the image contrast. Fig. 4 demonstrates the results of applying histogram equalization.

Fig. 4 demonstrates the results of histogram equalization, revealing that the processed image displays a more uniform distribution of grayscale values. This enhancement leads to improved brightness and more defined details. Additionally, Fig. 5 provides a sample image that has undergone histogram equalization, showcasing the resultant effects.

3.3. Cluster-based segmentation

Following the pre-processing, image pixels are clustered utilizing the GMM model. The process begins by initializing the maximum number of

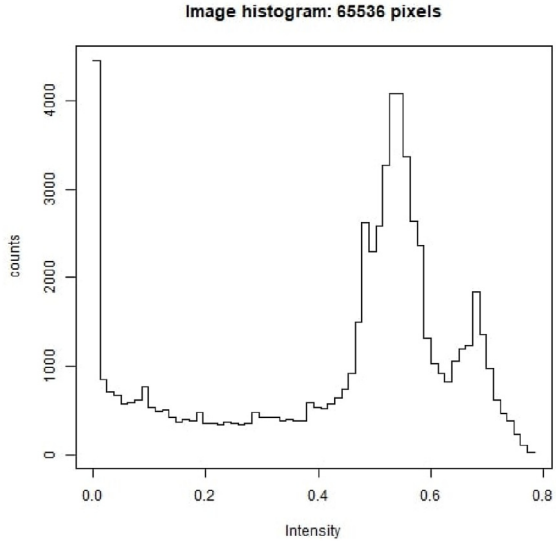

Fig. 3. Histogram of the original image.

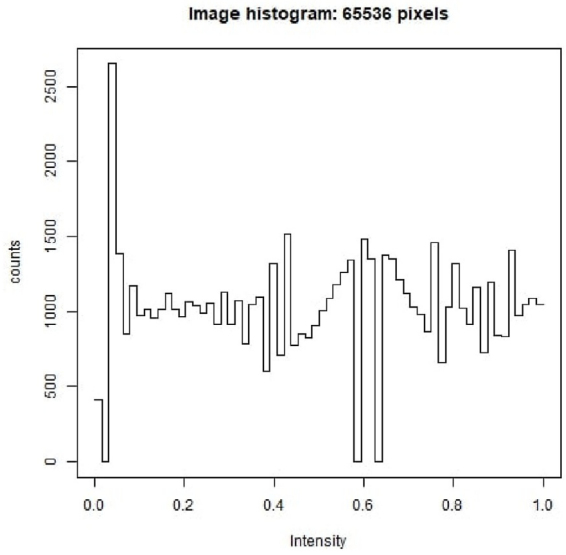

Fig. 4. Histogram equalization.

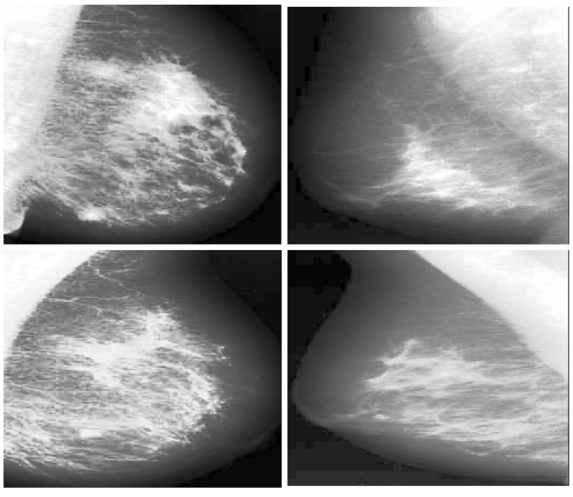

Fig. 5. Histogram equalization results.

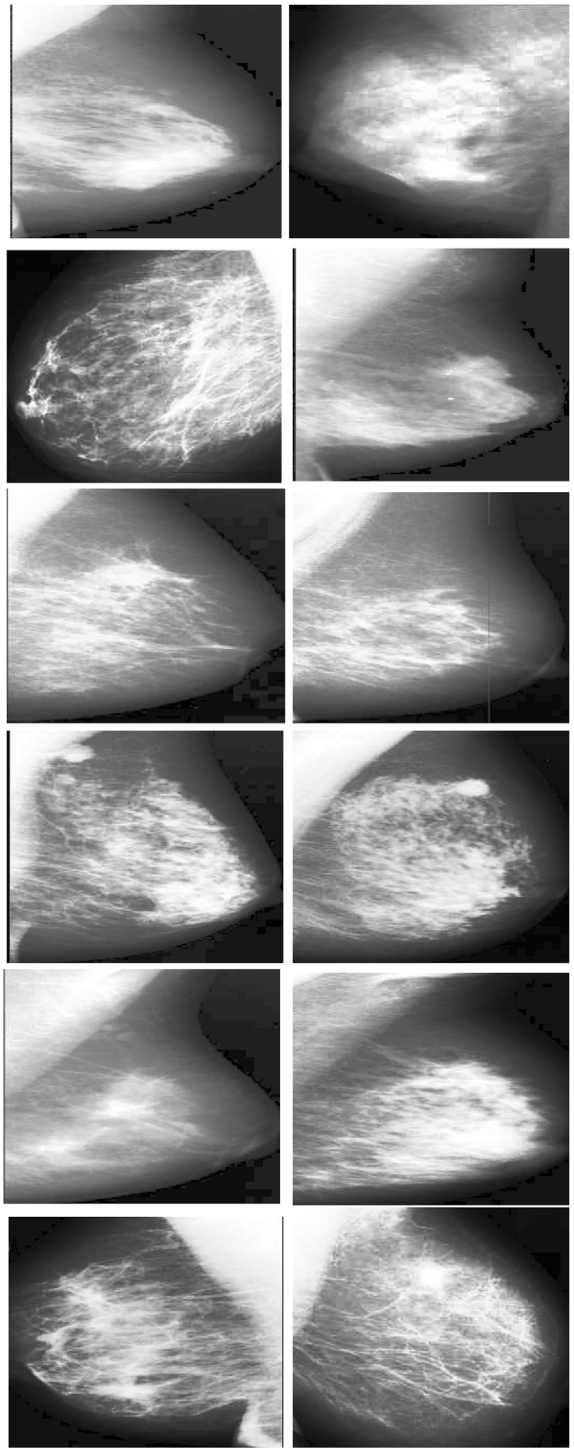

Fig. 5. (Continued).

clusters and then converting the initial clustering results into indicator variables. The parameters of the GMM are then optimized using the EM algorithm. The optimal cluster size is then determined by evaluating each model's BIC, where a higher

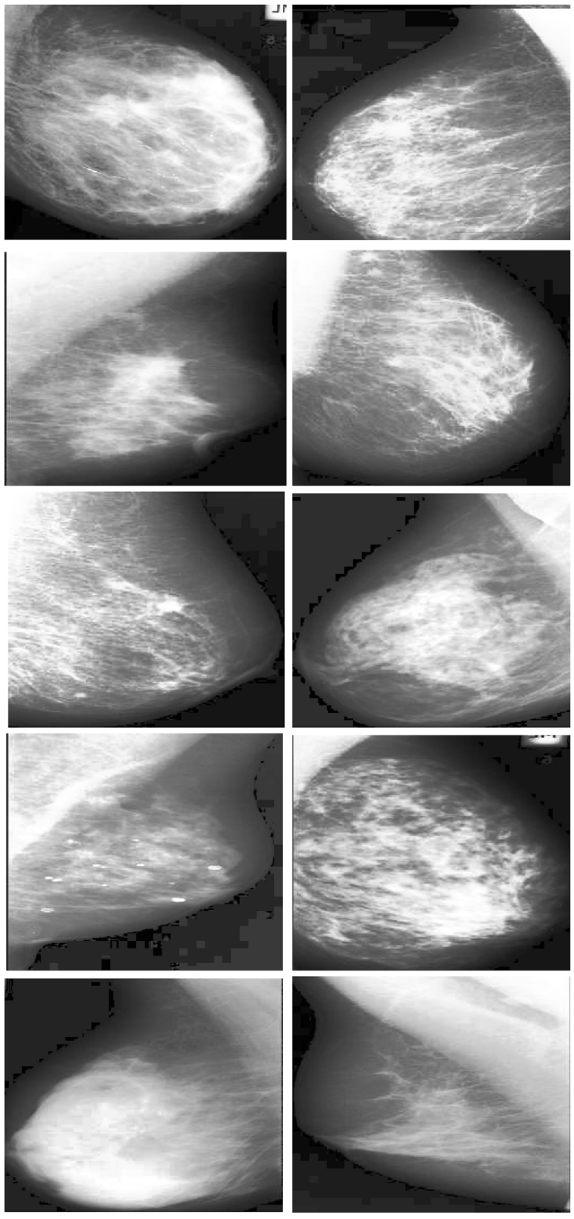

Fig. 5. (Continued).

BIC value indicates a better fit to the data. The results of the clustering process are illustrated in Fig. 6.

Fig. 6 illustrates the clustering outcomes using the GMM, demonstrating that the Variance-Equalized Independence (VEI) model with up to nine clusters yields the optimal performance for image segmentation. The model achieved a maximum BIC value of over 150.000. Consequently, during the ROI stage, which involves thresholding and morphological operations, the process employs the nine-cluster model derived from the VEI diagonal configuration for further analysis.

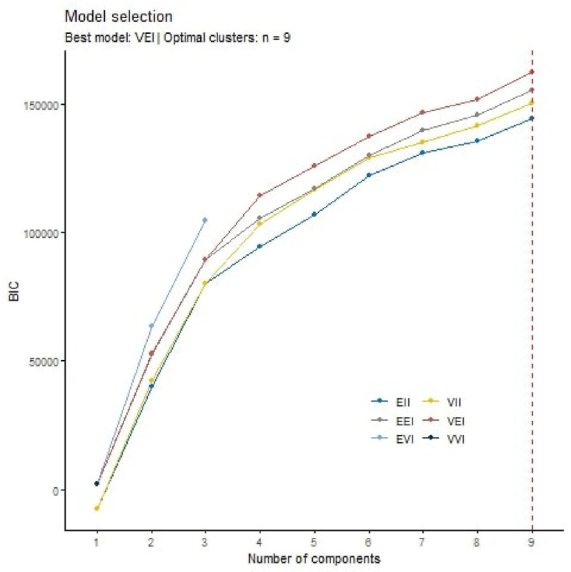

Fig. 6. Plot of GMM model.

3.4. Region of interest

The ROI stage encompasses both thresholding and morphological operations. Thresholding determines the optimal threshold value T_{opt} based on the maximum number of clusters m derived from the GMM model. Each image in the sample undergoes clustering, with the number of clusters capped at a maximum of nine. However, variability among images arises from their distinct parameters, such as $(\omega_i, \pi, \mu_j, \sigma_j)$, leading to differences in the threshold values applied to each image.

The T_{opt} value is computed as the average of the mean $\sum_{i=1}^{m}\mu_m$ during M-Step of the model, as calculated by Equation (6):

$$T_{opt} = \frac{1}{m}\sum_{i=1}^{m}\mu_m \tag{6}$$

After successfully determining the T_{opt} value, the next step involves extracting the identified cancerous regions through the opening operation—a morphological technique that preserves only the most prominent area corresponding to the breast. Subsequently, the fillHull function in R Studio is utilized to close any gaps in the binary image. The process concludes with a masking procedure that overlays the segmented breast image onto the original mammogram to produce the final segmented image.

The masking process poses a considerable challenge in the identification of the ROI. Errors that occur during this stage, especially when the segmented regions are overlaid onto the original images, can hinder the precise ROI identification. When segmentation does not accurately represent the underlying tissues, significant discrepancies may arise in detecting of cancerous areas, ultimately undermining the reliability of the diagnostic process. Fig. 7 shows a simplified depiction of the cancer image both before and after the masking process.

The detailed results of the images that have progressed through each stage of the ROI are presented in Fig. 8.

The results demonstrate that the EM-GMM algorithm effectively segmented twenty-six mammogram images into nine pre-defined clusters. This segmentation underscores the algorithm's ability to detect and delineate specific patterns within medical imaging data. Such segmentation is pivotal for enhancing the clarity and accuracy in identifying critical or anomalous regions in mammograms. By providing a more detailed understanding of each cluster, the process supports radiologists in pinpointing critical areas for early detection and

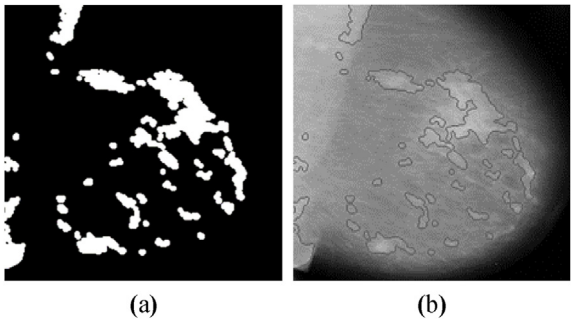

Fig. 7. ROI process, (a) morphology, (b) masking.

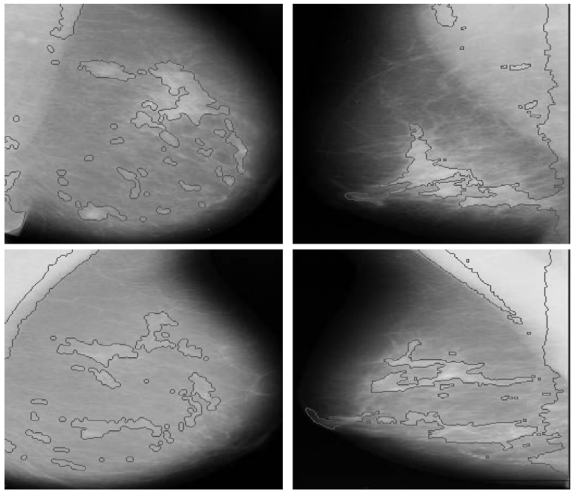

Fig. 8. Mammogram image segmentation results.

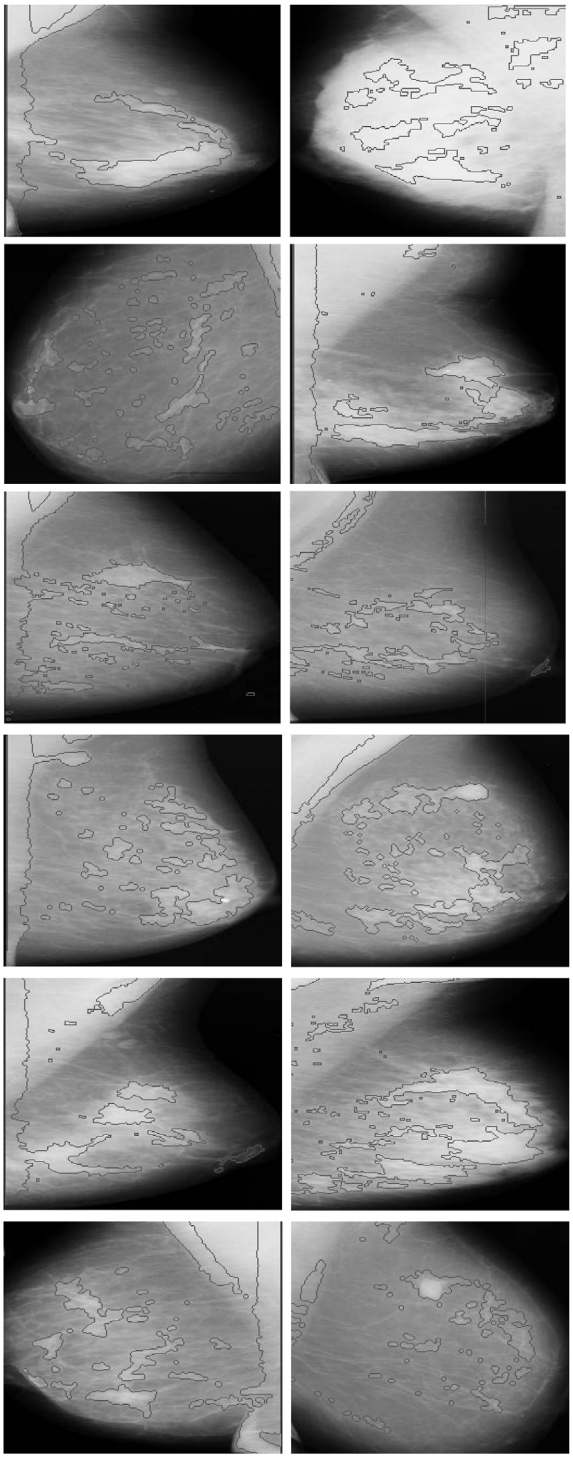

Fig. 8. Mammogram image segmentation results.

conducting thorough breast health evaluations. Furthermore, segmenting mammogram images into relevant clusters enhances the ability to differentiate and interpret the underlying patterns present in various regions of the images, thereby enabling a more comprehensive analysis of breast cancer data.

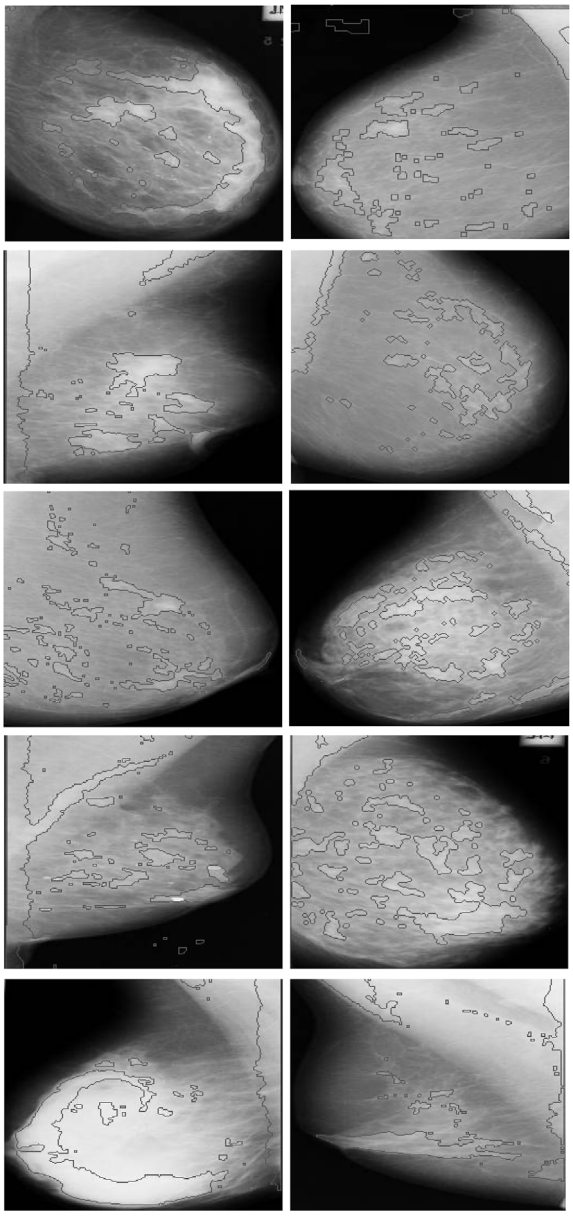

Fig. 8. (Continued)

3.5. Testing

The testing procedure involved assessing images with varying cluster numbers based on BIC values to identify the optimal segmentation for breast cancer mammogram data. Specifically, images were evaluated using 7, 8, and 9 clusters, respectively, which exhibited the highest BIC values throughout the segmentation procedure with EM-GMM, respectively. A comparative analysis of the segmentation outcomes is depicted in Fig. 9, highlighting the partitioning that best represents the distinct features of each image. This step is crucial to

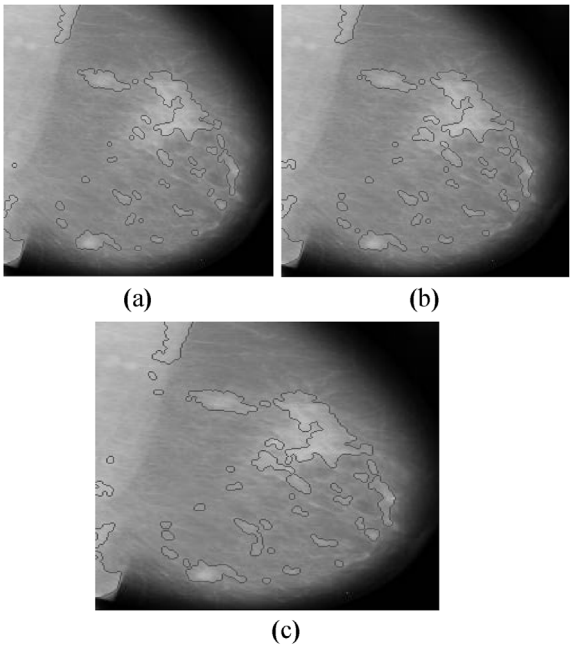

Fig. 9. Comparison of image segmentation with (a) 7 clusters, (b) 8 clusters, (c) 9 clusters.

ensure that the chosen number of clusters aligns with the objectives of the data analysis.

The test results indicate that image segmentation with nine clusters creates the highest segmentation accuracy. This conclusion is derived from a comparative analysis of segmentations with seven, eight, and nine clusters. As illustrated in Fig. 9, the region identified as potentially cancerous is progressively larger with an increasing number of clusters: smaller with seven clusters compared to eight and smaller with eight clusters compared to nine. Consequently, segmentation with nine clusters provides the most precise identification of suspected cancerous areas. This pattern is consistently observed across all segmented images.

3.6. Comparison with previous studies

In the field of breast cancer detection, numerous significant studies have applied various methodologies and objectives in image segmentation, each contributing to advancements in this crucial area. For instance, A. Arjmand, A. Farzamnia, R. Afrouzian, and S. Meshgini [25] employed an MRI-based approach that integrated K-Means with the CSO algorithm to detect tumour masses within in the RIDER breast dataset. Their approach demonstrated notable superiority over traditional methods like fuzzy C-means and standard K-means, though segmentation accuracy was not explicitly quantified.

Similarly, J. Dabass, M. Hanmandlu, and R. Vig [26] applied K-means clustering to digital mammograms from the Mini-Mias dataset, successfully segmenting images into different density regions. This study highlighted the effectiveness of K-means for breast cancer density segmentation, with accuracy ranging from 94.8 % to 98.75 % depending on the detected region, whether dense, fatty, or breast edges.

In another study, J. S. Isaac, K. Priya, M. J. Kumar, S. Kottu, V. Senthilkumar, and V. S. Ramakrishna [27] utilized machine learning techniques for breast cancer classification using ultrasonography, integrating K-means in the segmentation process. They achieved notable accuracy rates: 95.87 % with logistic regression, 97.14 % with random forest, and 93.33 % with K-nearest neighbour, demonstrating the effectiveness of their classification methods, though without a focus on clustering optimization.

F. G. Y. Ciklacandir, A. Ertaylan, U. Binzat, and A. Kut [28] also implemented K-means to breast ultrasonography for lesion detection, combining various filtering techniques such as median, Laplace, and Sobel filters. Their results, with accuracy ranging from 71.21 % to 77.08 %, revealed the significant impact of filtering on segmentation performance. However, the study broadly categorized lesions as benign or malignant without addresing on precise tumour area segmentation.

H. Lin and Z. Ji [30] applied a hybrid model combining K-means with a Self-Organizing Map (SOM) on the Wisconsin breast cancer dataset, achieving 92.10 % precision and 98.30 % recall. Their work focused on classifying tumours as benign or malignant, emphasizing classification performance over the enhancing clustering techniques.

K. Wisaeng [33] utilized a K-Means++ and CSO approach to multiple datasets (Mini-MIAS, DDSM, BCDR), achieving accuracy rates of 96.42 % for Mini-MIAS, 95.49 % for DDSM, and 96.92 % for BCDR. The study emphasized the importance of lesion tracking but did not optimize the number of clusters for malignancy identification.

Finally, L. Panigrahi, and R. R. Panigrahi [34] introduced an Ant Colony Optimization (ACO) approach alongside K-means for ultrasonography, achieving an accuracy of 91.66 %. Their approach enhanced traditional segmentation methods.

In contrast, the proposed research introduces an EM-GMM clustering methodology designed not only to segment breast cancer images but also to determine the optimal number of clusters for malignancy identification. By establishing that nine clusters offer the most accurate segmentation based on the BIC, this study improves upon prior methodologies. It integrates a robust probabilistic framework, thereby advancing breast

Table 2. Comparative analysis of research outcomes in mammographic image segmentation.

Researchers	Research outcomes
A. Arjmand, A. Farzamnia, R. Afrouzian, and S. Meshgini [25]	The results on the RIDER breast dataset clearly demonstrate that the proposed method outperforms other approaches, such as fuzzy c-means and simple k-means. Mean: 0.4506 Standard Deviation: 0.1843
J. Dabass, M. Hanmandlu, and R. Vig [26]	Accuracy ranges from 94.8 % to 98.75 % depending on the detected region, whether dense, fatty, or breast edges.
J. S. Isaac, K. Priya, M. J. Kumar, S. Kottu, V. Senthilkumar, and V. S. Ramakrishna [27]	Accuracy: 95.87 % (Logistic Regression) 97.14 % (Random Forest) 93.33 % (K-Nearest Neighbor)
F. G. Y. Ciklacandir, A. Ertaylan, U. Binzat, A. Kut [28]	Accuracy: 76.56 % (Tanpa filter) 77.08 % (Median) 71.21 % (Laplace) 71.25 % (Sobel)
P. P. Golagani, S. K. Beebi, T. S. Mahalakshmi [29]	The model encompasses a total of eight pathways.
H. Lin and Z. Ji [30]	Precision: 92.10 % Recall: 98.30 % F1-score: 95.00 %
D. N. Ouedraogo [31]	Best model (Adaboost classifier): 97.90 % Best model cluster: 1
S. H. Abdulla, A. M. Sagheer, H. Veisi [32]	Accuracy: 98.2 %. Sensitivity: 97.05 %. Specificity: 98.52 %.
K. Wisaeng [33]	Accuracy: 96.42 % (Mini−MIAS) 95.49 % (DDSM) 96.92 % (BCDR)
L. Panigrahi, R. R. Panigrahi [34]	Accuracy: 91.66 %
A. S. Begum, K. Bhuvaneshwari, K. J. Sri, M. Divya, and R. M. Prakash [35]	The qualitative analysis indicates that the FCM segmentation gives good accuracy and indication of the disease.
I. Khoulqi, N. Idrissi, and M. Sarfraz [37]	DICE: 88.23 % SSIM: 89.58 %
Proposed study	The optimal number of clusters for the segmentation or estimation of breast cancer regions is identified as nine clusters. This determination is based on the BIC calculations. An evaluation was conducted by testing the images with seven and eight clusters.

cancer detection through enhanced clustering techniques.

In summary, a comparative analysis of the findings of prior studies and the results obtained in this research is provided in Table 2.

4. Conclusion

This study explores the application of the EM-GMM algorithm to accurately identifying the precise location and dimensions of breast cancer. The results show that incorporating pre-processing steps, histogram equalization, and cluster-based segmentation with the GMM model produces reliable outcomes for analysing mammogram images. This method shows considerable promise in improving early detection and comprehensive evaluation of breast health. Enhanced segmentation accuracy allows radiologists to more precisely locate critical areas, thereby increasing diagnostic accuracy and facilitating more effective treatment planning. The research results significant enhance the effectiveness of diagnosing and treating breast cancer, potentially leading to a reduction in both mortality and morbidity associated with the condition.

Ethics information

This research did not include any human subjects or animals, and as such, it was not necessary to obtain ethical approval, in accordance with both institutional and national regulations and guidelines.

Funding

This study was not supported by any specific grants or financial contributions from funding bodies within the public, commercial, or nonprofit sectors.

Acknowledgements

We want to express our gratitude to all those who contributed to the research and preparation of this work.

References

[1] A. Bleyer, H.G. Welch, Effect of three decades of screening mammography on breast cancer incidence, N. Engl. J. Med. 367 (2012) 1998–2005, https://doi.org/10.1056/nejmoa1206809.
[2] B. Hela, M. Hela, H. Kamel, B. Sana, M. Najla, Breast cancer detection: a review on mammograms analysis techniques, in: 2013 10th International Multi-Conference on Systems, Signals and Devices, SSD 2013, Hammamet, Tunisia, 2013, pp. 1–6, https://doi.org/10.1109/SSD.2013.6563999.
[3] M. Løberg, M.L. Lousdal, M. Bretthauer, M. Kalager, Benefits and harms of mammography screening, Breast Cancer Res. 17 (2015) 1–12, https://doi.org/10.1186/s13058-015-0525-z.
[4] L. Wang, Early diagnosis of breast cancer, Sensors 17 (2017) 1–20, https://doi.org/10.3390/s17071572.
[5] O. Ginsburg, C. Yip, A. Books, A. Cabanes, M. Caleffi, J.A.D. Yataco, B. Gyawali, V. McCormack, M.M. de Anderson, R. Mehrotra, A. Mohar, R. Murillo, L.E. Pace, E.D. Paskett, A. Romanoff, A.F. Rositch, J.R. Scheel, M. Schneidman, K. Unger-Saldana, V. Vanderpuye, T. Wu, S. Yuma, A. Dvaladze, C. Duggan, B.O. Anderson, Breast cancer early detection: a phased approach to implementation, Cancer 126 (2020) 2379–2393, https://doi.org/10.1002/cncr.32887.
[6] E. Michael, H. Ma, H. Li, F. Kulwa, J. Li, Breast cancer segmentation methods: current status and future potentials, BioMed Res. Int. 2021 (2021) 1–29, https://doi.org/10.1155/2021/9962109.
[7] R.C. Fitzgerald, A.C. Antoniou, L. Fruk, N. Rosenfeld, The future of early cancer detection, Nat. Med. 28 (2022) 666–677, https://doi.org/10.1038/s41591-022-01746-x.
[8] M. Alsammak, M. Khattabi, Breast cancer screening a literature review, Int. J. Community Med. Public Heal. 10 (2023) 4473–4479, https://doi.org/10.18203/2394-6040.ijcmph20233497.
[9] N.M. ud Din, R.A. Dar, M. Rasool, A. Assad, Breast cancer detection using deep learning: datasets, methods, and challenges ahead, Comput. Biol. Med. 149 (2022) 106073, https://doi.org/10.1016/j.compbiomed.2022.106073.
[10] J.V. Fiorica, Breast cancer screening, mammography, and other modalities, Clin, Obstet Gynecol 59 (2016) 688–709, https://doi.org/10.1097/GRF.0000000000000246.
[11] M. Abdel-Nasser, F.P. Solsona, D. Puig, Pectoral muscle segmentation in tomosynthesis images using geometry information and grey wolf optimizer, in: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, SciTePress, Valletta, Malta. 2020, pp. 829–836, https://doi.org/10.5220/0009156408290836.
[12] A. Rodriguez-Ruiz, J. Teuwen, K. Chung, N. Karssemeijer, M. Chevalier, A. Gubern-Merida, I. Sechopoulos, Pectoral muscle segmentation in breast tomosynthesis with deep learning, in: Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis, Texas, United States, 2018, p. 105752J, https://doi.org/10.1117/12.2292920.
[13] S.Y. Guraya, Breast cancer screening: implications and clinical perspectives, J. Taibah Univ. Med. Sci. 3 (2008) 67–82, https://doi.org/10.1016/s1658-3612(08)70056-8.
[14] M.T. Tirona, Breast cancer screening update, Am. Fam. Physician. 87 (2013) 274–278. https://www.aafp.org/pubs/afp/issues/2013/0215/p274.pdf.
[15] A. Nedra, M. Shoaib, S. Gattoufi, Detection and classification of the breast abnormalities in digital mammograms via linear support vector machine, in: Middle East Conference on Biomedical Engineering, IEEE, Tunis, Tunisia. 2018, pp. 141–146, https://doi.org/10.1109/MECBME.2018.8402422.
[16] A.H. Abbas, A.A. Kareem, M.Y. Kamil, Breast cancer image segmentation using morphological operations, Int. J. Electron. Commun. Eng. Technol. 6 (2015) 8–14. https://iaeme.com/Home/article_id/40120150604002.
[17] J. Dheeba, S.T. Selvi, A swarm optimized neural network system for classification of microcalcification in mammograms, J. Med. Syst. 36 (2012) 3051–3061, https://doi.org/10.1007/s10916-011-9781-3.
[18] S. Sasikala, M. Ezhilarasi, P. Sudharsan, C.L. Yashwanthi Sivakumaric, Performance analysis of various segmentation techniques in breast mammogram images, in: Proceedings - 2014 International Conference on Intelligent Computing Applications, IEEE, Coimbatore, India. 2014, pp. 228–232, https://doi.org/10.1109/ICICA.2014.56.
[19] L.J. Grimm, C.S. Avery, E. Hendrick, J.A. Baker, Benefits and risks of mammography screening in women ages 40 to 49 years, J. Prim. Care Community Heal. 13 (2022) 1–6, https://doi.org/10.1177/21501327211058322.
[20] C. De Jesus, T.W. Moseley, V. Diaz, V. Vishwanath, S. Jean, A. Elhatw, H.R.F.D. Pria, H.L. Chung, M.S. Guirguis, M.M. Patel, The benefits of screening mammography, Curr. Breast Cancer Rep. 15 (2023) 103–107, https://doi.org/10.1007/s12609-023-00479-1.
[21] S.O. Khanbari, A.S.M. Haider, Enhanced mammography image for breast cancer detection using LC-CLAHE technique, Univ. Aden J. Nat. Appl. Sci. 24 (2020) 143–154, https://doi.org/10.47372/uajnas.2020.n1.a12.
[22] S.M. Badawy, A.A. Hefnawy, H.E. Zidan, M.T. GadAllah, Breast cancer detection with mammogram segmentation: a qualitative study, Int. J. Adv. Comput. Sci. Appl. 8 (2017) 117–120, https://doi.org/10.14569/ijacsa.2017.081016.
[23] N.S. Hassan, A.M. Abdulazeez, D.Q. Zeebaree, D.A. Hasan, Medical images breast cancer segmentation based on k-means clustering algorithm: a review, Asian J. Res. Comput. Sci. (2021) 23–38, https://doi.org/10.9734/ajrcos/2021/v9i130212.
[24] Y.G. Jung, M.S. Kang, J. Heo, Clustering performance comparison using k-means and expectation maximization algorithms, Biotechnol. Biotechnol. Equip. 28 (2014) S44–S48, https://doi.org/10.1080/13102818.2014.949045.
[25] A. Arjmand, S. Meshgini, R. Afrouzian, A. Farzamnia, Breast tumour segmentation using k-means clustering and cuckoo search optimization, in: 2019 9th International Conference on Computer and Knowledge Engineering, IEEE, Mashhad, Iran. 2019, pp. 305–308, https://doi.org/10.1109/ICCKE48569.2019.8964794.
[26] J. Dabass, M. Hanmandlu, R. Vig, Segmentation of breast density using K-means clustering algorithm, in: U. Batra, N. Roy, B. Panda, eds., Data Science and Analytics, Springer, Singapore. 2020, pp. 305–315, https://doi.org/10.1007/978-981-15-5830-6_26.
[27] K. Priya, V. Senthilkumar, J.S. Isaac, S. Kottu, V.S. Ramakrishna, M.J. Kumar, Breast cancer segmentation by k-means and classification by machine learning, in: International Conference on Automation, Computing and Renewable Systems, IEEE, Pudukkottai, India. 2022, pp. 651–656, https://doi.org/10.1109/ICACRS55517.2022.10029301.
[28] F.G.Y. Ciklacandir, A. Ertaylan, U. Binzat, A. Kut, Lesion detection from the ultrasound images using k-means algorithm, in: TIPTEKNO 2019 - Tip Teknolojileri Kongresi, IEEE, Izmir, Turkey. 2019, pp. 68–71, https://doi.org/10.1109/TIPTEKNO.2019.8895050.
[29] P.P. Golagani, S.K. Beebi, T.S. Mahalakshmi, Using K-means clustering algorithm with Python programming for predicting breast cancer, in: J. Fiaidhi, D. Bhattacharyya, N. Rao, eds., Smart Technologies in Data Science and Communication, Springer, Singapore. 2020, pp. 163–171, https://doi.org/10.1007/978-981-15-2407-3_21.
[30] H. Lin, Z. Ji, Breast cancer prediction based on k-means and som hybrid algorithm, J. Phys. Conf. Ser. 1624 (2020) 1–6, https://doi.org/10.1088/1742-6596/1624/4/042012.

[31] D.N. Ouedraogo, Interpretable machine learning model selection for breast cancer diagnosis based on k-means clustering, Appl. Med. Informatics Orig. Res. 43 (2021) 91–102. https://ami.info.umfcluj.ro/index.php/AMI/article/view/817/816.
[32] S.H. Abdulla, A.M. Sagheer, H. Veisi, Breast cancer segmentation using k-means clustering and optimized region-growing technique, Bull. Electr. Eng. Informatics. 11 (2022) 158–167, https://doi.org/10.11591/eei.v11i1.3458.
[33] K. Wisaeng, Breast cancer detection in mammogram images using k−means++ clustering based on cuckoo search optimization, Diagnostics 12 (2022) 1–21, https://doi.org/10.3390/diagnostics12123088.
[34] L. Panigrahi, R.R. Panigrahi, An enhancement in K-means algorithm for automatic ultrasound image segmentation, in: B.K. Singh, G. Sinha, R. Pandey, eds., Biomedical Engineering Science and Technology, Springer, Cham. 2024, pp. 1–8, https://doi.org/10.1007/978-3-031-54547-4_1.
[35] R.M. Prakash, K. Bhuvaneshwari, M. Divya, K.J. Sri, A.S. Begum, Segmentation of thermal infrared breast images using k-means, FCM and EM algorithms for breast cancer detection, in: Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, IEEE, Coimbatore, India. 2017, pp. 1–4, https://doi.org/10.1109/ICIIECS.2017.8276142.
[36] A.H. Yurttakal, H. Erbay, T. İkizceli, S. Karaçavuş, G. Çinarer, A comparative study on segmentation and classification in breast MRI imaging, IIOAB J. 9 (2018) 23–33. http://www.iioab.org/article/IIOABJ_9.5_23-33.pdf.
[37] I. Khoulqi, N. Idrissi, M. Sarfraz, Segmentation of pectoral muscle in mammogram images using Gaussian mixture model-expectation maximization, in: I. Management Association, eds., Research Anthology on Medical Informatics in Breast and Cervical Cancer, IGI Global, 2023, pp. 722–738, https://doi.org/10.4018/978-1-6684-7136-4.ch038.
[38] Q. Wang, Y. Yuan, Learning to resize image, Neurocomputing 131 (2014) 357–367, https://doi.org/10.1016/j.neucom.2013.10.007.
[39] N. Rathee, S. Pahal, D. Sheoran, Evaluating the uncertainty of classification due to image resizing techniques for satellite image classification, Mapan - J. Metrol. Soc. India 36 (2021) 243–251, https://doi.org/10.1007/s12647-021-00456-y.
[40] S. Saponara, A. Elhanashi, Impact of image resizing on deep learning detectors for training time and model performance, in: S. Saponara, A. De Gloria, eds., Applications in Electronics Pervading Industry, Springer, Cham. 2022, pp. 10–17, https://doi.org/10.1007/978-3-030-95498-7_2.
[41] P. Härtinger, C. Steger, Adaptive histogram equalization in constant time, J. Real-Time Image Proc. 21 (2024) 1–9, https://doi.org/10.1007/s11554-024-01465-1.
[42] S. Zeng, R. Huang, Z. Kang, N. Sang, Image segmentation using spectral clustering of Gaussian mixture models, Neurocomputing 144 (2014) 346–356, https://doi.org/10.1016/j.neucom.2014.04.037.
[43] A. Bojchevski, Y. Matkovic, S. Günnemann, Robust spectral clustering for noisy data, in: KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Digital Library, Canada. 2017, pp. 737–746, https://doi.org/10.1145/3097983.3098156.
[44] Y. Kim, H. Do, S.B. Kim, Outer-Points shaver: robust graph-based clustering via node cutting, Pattern Recognit. 97 (2020) 1–13, https://doi.org/10.1016/j.patcog.2019.107001.
[45] L.T. Li, Z.Y. Xiong, Q.Z. Dai, Y.F. Zha, Y.F. Zhang, J.P. Dan, A novel graph-based clustering method using noise cutting, Inf. Syst. 91 (2020) 1–14, https://doi.org/10.1016/j.is.2020.101504.
[46] P.R. Srivastava, P. Sarkar, G.A. Hanasusanto, A robust spectral clustering algorithm for sub-Gaussian mixture models with outliers, Oper. Res. 71 (2023) 224–244, https://doi.org/10.1287/opre.2022.2317.
[47] J.H. Wolfe, Object Cluster Analysis of Social Areas, Doctoral Dissertation, University of California. 1963.
[48] D.N. Geary, Mixture models: inference and applications to clustering, in: G.J. McLachlan, K.E. Basford, eds., J. R. Stat. Soc. Ser. A (Statistics Soc.) 152 (1989) 126–127, https://doi.org/10.2307/2982840.
[49] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1977) 1–22, https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
[50] M. Ram, A probabilistic approach to reconfigurable interactive manufacturing and coil winding for industry 4.0, in: S. Michieletto, F. Stival, E. Pagello, eds., Advances in Mathematics for Industry 4.0, Academic Press, United States. 2020, pp. 61–93, https://doi.org/10.1016/B978-0-12-818906-1.00003-6.
[51] S.R. Kasa, V. Rajan, Avoiding inferior clusterings with misspecified Gaussian Mixture Models, Sci. Rep. 13 (2023) 1–13, https://doi.org/10.1038/s41598-023-44608-3.
[52] W. Jannah, D.R.S. Saputro, Parameter estimation of Gaussian mixture models (GMM) with expectation maximization (EM) algorithm, in: AIP Conference Proceedings, AIP Publishing, Indonesia. 2022 040002, https://doi.org/10.1063/5.0117119.
[53] N. Sammaknejad, Y. Zhao, B. Huang, A review of the Expectation Maximization algorithm in data-driven process identification, J. Process Control 73 (2019) 123–136, https://doi.org/10.1016/j.jprocont.2018.12.010.
[54] T. Shaik, X. Tao, L. Li, N. Higgins, R. Gururajan, X. Zhou, J. Yong, Clustered FedStack: intermediate global models with bayesian information criterion, Pattern Recognit. Lett. 177 (2024) 121–127, https://doi.org/10.1016/j.patrec.2023.12.004.
[55] M.S. Hossain, G.M. Shahriar, M.M.M. Syeed, M.F. Uddin, M. Hasan, S. Shivan, S. Advani, Region of interest (ROI) selection using vision transformer for automatic analysis using whole slide images, Sci. Rep. 13 (2023) 1–14, https://doi.org/10.1038/s41598-023-38109-6.
[56] R.C. Gonzalez, R.E. Woods, Digital Image Processing, second ed., Prentice Hall, New Jersey. 2002.
[57] J.C. Bezdek, L.O. Hall, L.P. Clarke, Review of MR image segmentation techniques using pattern recognition, Med. Phys. 20 (1993) 1033–1048, https://doi.org/10.1118/1.597000.
[58] S. Mondal, P. Bours, A study on continuous authentication using a combination of keystroke and mouse biometrics, Neurocomputing 230 (2017) 1–22, https://doi.org/10.1016/j.neucom.2016.11.031.
[59] A. Shukla, S. Kanungo, An efficient clustering-based segmentation approach for biometric image, Recent Adv. Comput. Sci. Commun. 14 (2020) 803–819, https://doi.org/10.2174/2666255813666200219153105.
[60] S. Jardim, J. António, C. Mora, Image thresholding approaches for medical image segmentation-short literature review, Procedia Comput. Sci. 219 (2023) 1485–1492, https://doi.org/10.1016/j.procs.2023.01.439.
[61] H.A. Saeed, S. Hamad, A.T. Hussain, Analysis the digital images by using morphology operators, Indones. J. Electr. Eng. Comput. Sci. 24 (2021) 1654–1662, https://doi.org/10.11591/ijeecs.v24.i3.pp1654-1662.