IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 38, No. 2, April, 2024, pp. 439-448
Journal homepage: http://iieta.org/journals/ria

Mammographic Masses Descriptor for Breast Cancer Classification and Automatic Diagnosis

Mohammed EL Amine Yermes[1*], Mohammed Debakla[1], Khalifa Djemal[2]

[1] Computer Science Department, University of Mustapha Stambouli, Mascara 29000, Algeria
[2] IBISC Laboratory, Evry Val d'Essone University, Paris 91020, France

Corresponding Author Email: amine.yermes@univ-mascara.dz

https://doi.org/10.18280/ria.380207

Received: 6 August 2023
Revised: 20 January 2024
Accepted: 19 February 2024
Available online: 24 April 2024

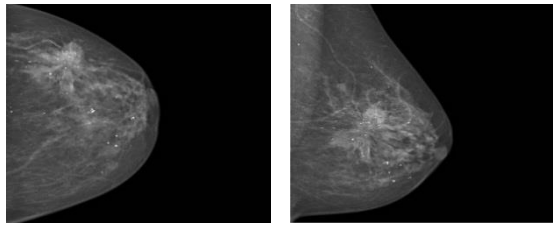
Keywords:
breast cancer, spiculated masses, triangle-area representation, computer aided diagnosis, mass description, mammography

ABSTRACT

An automatic breast cancer diagnosis is a challenging task because breast masses have a random appearance and vary in size and shape. A descriptor is an algorithm that quantifies elementary characteristics such as color, texture, contour, or shape. In digital mammography, numerous descriptors have been employed to differentiate between benign and malignant tumor patterns, but automatic diagnosis remains a difficult function. In this paper, we proposed a novel approach based on local features to describe masses in mammograms via Polygon Approximation Triangle-Area Representation (PATAR). As the degree of spiculation in masse defines their level of malignancy, the strength of our approach lies in its ability to isolate and measure spiculations in breast masses. PATAR is a robust image descriptor composed of two steps: polygon approximation and triangle-area representation. Firstly, we applied a polygon approximation to the masses to raise the most critical spiculations and lobulations. Then, by browsing the points of the polygon, calculate the triangle's area formed by the vertices of the polygon, ears, and mouths. The extracted characteristics describe the shape and show the severity of spiculations with high precision. Digital mammography CBIS-DDSM is used to evaluate the method with a Fuzzy C-Means classifier, Support Vector Machines (SVM), and Random Forest (RF). The Random Forest classifier achieved the best performance, reaching 97,94%. The proposed method provides a fully automated diagnosis with the best accuracy and invariance to scale and rotation.

1. INTRODUCTION

Breast cancer accounts for one in four cancer cases in women globally, making it the most often diagnosed malignancy. The reported 2.3 million new cases suggest that approximately one out of every eight cancer diagnoses in the year 2020 was related to breast cancer [1]. Breast cancer is a severe worldwide health concern, and its prevalence differs geographically, with significant rates observed in third-world countries. Breast cancer development is linked to several risk factors, which may be categorized primarily as non-modifiable and modifiable factors. Non-modifiable risk variables, including gender, age, and genetics. Modifiable risk factors are Lifestyle Factors, Hormone and Replacement Therapy (HRT) and Reproductive Factors. Diagnosing breast cancer involves a combination of screening like mammography and ultrasound, clinical examination, and diagnostic tests such as biopsy. Early detection plays a crucial role in the successful treatment of breast cancer, mainly when the tumor is smaller and has not metastasized, increasing the likelihood of successful outcomes. Automated diagnostic systems, particularly those based on artificial intelligence, have proven to enhance sensitivity and specificity compared to traditional approaches. These progressions play a role in achieving more precise and dependable diagnoses of breast cancer. Automated diagnosis can help radiologists rapidly categorizing and prioritizing cases, potentially reducing the time to diagnosis and treatment. An efficient system of automatic diagnosis can reduce significant deaths and improve the lives of cancer patients. Mammography is the most effective radiological technique and is widely used for the early detection and diagnosis of breast cancer [2]. A mammogram is a special electronic detector based on an X-ray that has been used since 1913 to examine breasts [3]. Mammograms provide a digital viewing of breast images with two standard views: bilateral craniocaudal (CC), shown in Figure 1(a), and mediolateral oblique (MLO), shown in Figure 1(b).



(a) (b)

Figure 1. Two standard views of mammographic image**:** (a) bilateral craniocaudal (CC) view and (b) mediolateral oblique (MLO) view

In the late 1980s, the American College of Radiology (ACR) launched the Breast Imaging-Reporting and Data System (BI-RADS) to provide radiologists with standardized terminology [4]. The BI-RADS system provides a common language for radiologists and healthcare providers to communicate about breast imaging results, facilitating consistent interpretation and appropriate patient management [5]. BI-RADS categorizes breast imaging findings into several assessment categories, ranging from 0 to 6. These categories help convey the level of suspicion for malignancy and guide subsequent actions. In the context of automatic breast cancer diagnosis, the BI-RADS system plays a significant role in training and validating machine learning algorithms. These algorithms can be trained on large datasets labeled with BI-RADS categories to learn patterns associated with different levels of suspicion. BI-RADS covers the major abnormalities of breast cancer, namely masses and microcalcifications. Shape and contour are the most significant features that indicate whether a mass is benign or malignant. Masse's shape may be round, oval, lobular, or irregular. Generally, benign masses tend to be round, oval and circumscribed. On the other hand, malignant masses suggest a greater likelihood of irregular shape and spiculated contour [5].

In the last few decades, due to the evolution of mammography, automatic diagnosis has occupied an important place in the domain of breast cancer recognition. Computer-aided diagnosis (CAD) has become an interdisciplinary field, including image processing, machine learning, computer vision, mathematics, physics, and statistics. This combination creates advanced computer tools that help radiologists in the interpretation process. Computer Aided Diagnosis or CADx systems ensure, in most cases, a fully automated double reading of mammographic images for radiologists. CAD (Computer-Aided Diagnosis) systems aid radiologists by assisting in detecting abnormalities, but they cannot replace radiologists. Human expertise and judgment are essential for accurate diagnosis and patient care. These elements are integrated into computerized methods designed to support radiologists in their medical decision-making processes. The primary goal of diagnostic systems is to enhance the precision of a radiologist [6]. CADx systems comprise three basic steps, segmentation, description, and, classification. This evolution in CAD systems is due to shape representation and characterization methods applied to digital mammographic masses [7]. Despite using different descriptors based on various feature extraction techniques, the problem of describing spiculated masses remains difficult. The blurred margins and random shapes of masses make the feature extraction and classification process complicated. To overcome this problem, robust descriptors based on significant features are necessary.

This paper proposes a novel approach based on the Triangle-area representation with polygon approximation (PATAR) for detecting and characterizing convexities and concavities in masses. The primary purpose of this paper is to design an advanced, fully automated method for diagnosing breast cancer in digital mammograms. Our interest was to develop a highly accurate descriptor in detecting lobulations and spiculations of masses. The precision in measuring mass features significantly affects the classification process [8]. To overcome the problem of spiculation, lobulation detection, and quantification, a polygon approximation is applied to the masses to raise the most critical convexities and ignore minor variations in contours. The triangles made by the polygon approximation step are measured with the Triangle-area representation (TAR) algorithm. TAR signature allows the PATAR algorithm to isolate and reasonably estimate the spiculated parts of masses. With this descriptor, the scale, rotation and translation invariant are ensured.

The remainder of this paper is organized as follows. Section 2 presents some of the recent methods and descriptors used in CADx systems. In section 3, our approach is detailed, including feature extraction and classification. Section 4 presents the classification process. Experimental results and comparison with some previous works are detailed in section 5. Section 6 concludes the paper and gives suggestions for future research.

2. RELATED WORK

Research on breast cancer classification involves various methodologies such as machine learning, deep learning, and image analysis. Studies focus on improving diagnostic accuracy, identifying biomarkers, and predicting cancer subtypes. Diverse datasets and innovative algorithms contribute to the ongoing advancements in breast cancer detection, aiding early diagnosis and personalized treatment. These tools also aid radiologists and technicians in quicker and more effective identification of breast cancer through mammography.

Guliato et al. [9] proposed a descriptor that employs polygons for identifying key characteristics of spicules, including recognition points and features, in addition to the previously mentioned aspects of spiculation. Then, the turning function is applied to the polygon to calculate the spiculation index. An accuracy of 93% was obtained on 111 images.

Xie et al. [10] built a model using the level set model for the detection and segmentation of masses, then feature extraction was done with multidimensional feature vectors covering gray level features combined with textural features, and feature selection is achieved by SVM and extreme learning machine (ELM) since not all feature vectors improve the classification. MIAS and DDSM datasets are used. The accuracy obtained is 96.02%. Beheshti et al. [11] Applied the fractal method as a region of interest (ROI); the fractal technique was based on discriminating lesions from the background tissues with a low mean square error of 0.21%. Then, evaluation for malignancy is done by defining new fractal features based on extracting asymmetric information from the lesion. Using 168 images they get an accuracy of 87.81%.

Souza et al. [12] suggested a method using section area, convolutions, and descriptors of shape distribution. The method utilizes a set of shape descriptors D1dist, D2dist, and D3dist, a modified descriptor of D1, D2, and D3 conceived by Osada et al. [13]. They proposed three methods, D1dist, D1dist, and D3dist, which are determined by selecting a group of random pixels on the surface of the mass. Section area and convolutions are also generated by dividing the object's surface in the x-y plane along the z-axis. For each descriptor, they calculate the standard deviation and mean of the shapes, resulting in 12 features. They use an SVM classifier, obtaining an accuracy of 92.15%.

Ribli et al. [14] Used a Faster Region-based Convolutional Neural Network (Faster R-CNN) to make an automatic system that recognizes and classifies lesions on a digital mammography. Faster R-CNN builds upon a convolutional neural network and includes extra components to identify,

localize, and classify items in an image. The base CNN utilized in this model was a VGG16 (Visual Geometry Group) network, which means a 16-layer deep CNN. In the last layer, two objects are detected in each image: benign or malignant lesions. They achieved a classification accuracy of about 85%. Goudarzi et al. [15] extract compactness, entropy, mean, and smoothness from mini-MIAS database images. Then, a fuzzy classifier with the look-ep method is performed, getting an accuracy of 89.37%.

Sun et al. [16] modified the Convolutional Neural Network (CNN) architecture to extract significant features from multiple views of MLO and CC to use the additional information from multiple mammography views. The model integrates a multi-view convolutional neural subnetwork (MVCNN) and a multi-dilated convolutional neural subnetwork (MDCNN) to extract features. They obtain an accuracy of 82.02% with the SVM classifier.

Vijayarajeswari et al. [17] use the Hough transform to separate particular shapes in an image with an SVM classifier. The image is processed with the Hough transform, similar to a random transform. This technique is primarily employed to detect diverse shapes and straight lines. The Hough transform exhibits tolerance towards gaps, noise, and occlusion in mammograms. The effective extraction and distinct separation of features are essential. Hough transform calculates An accumulator for each edge point (x, y). They used 95 images from the MIAS database, getting an accuracy of 94%. Pezeshki et al. [18] extract spiculated pixels of a tumor with homogenous intensity. To specify the correspondence of pixels in the same direction to indicate a spiculated part of the mass, they compute the summation of dissimilarities between the central pixel and its adjacent pixels across all symmetric orthogonal orientations. A minimal difference in each direction signifies pixel similarity and suggests the presence of spiculations. They considered the summation calculated in the previous step to extract spiculated components in all directions. The accuracy obtained with this method was 92.33%.

De Brito Silva et al. [19] developed a descriptor based on maps representing geometric and topological features and the distribution of shapes. Two spatial feature maps, namely the distance map and surface map, are computed for each mammography image. These maps describe mass geometry and topology. To compute similarity measures, they generate two spatial feature maps, the distance map and the surface map. These maps save features along with their corresponding spatial information, representing the geometry and topology of the mass. Additionally, shape descriptors built on distance histograms are employed to calculate distances, area, and angles, contributing to the characterization of mass shape. The best accuracy obtained was 93.70% using 794 images.

Arora et al. [20] Proposed a deep ensemble transfer learning for automatic feature extraction. A feature extraction technique based on a deep ensemble was introduced using a neural network (NN) classifier to evaluate the discriminative capability of the extracted features. This approach is implemented on pre-processed Region of Interest (ROI) image patches from the Curated Breast Imaging Subset Digital Database for Screening Mammography (CBIS-DDSM) dataset, generating features corresponding to each deep sub-architecture. An optimized feature vector is created by normalizing the utilized features, and this vector is subsequently input into the NN classifier for the final classification. Using SVM, the method achieves an accuracy of 88% with an Area Under the Curve (AUC) of 0.88.

Zhang et al. [21]. Develop an approach named BDR-CNN-GCN; this method integrates two advanced neural networks: (a) a graph convolutional network (GCN) and (b) a convolutional neural network (CNN). Initially, they employed a standard 8-layer CNN and incorporated two enhancement techniques: (a) batch normalization (BN) and (b) dropout (DO). Subsequently, they replaced traditional max pooling with rank-based stochastic pooling (RSP), creating BDR-CNN, a fusion of CNN, BN, DO, and RSP. BDR-CNN was then hybridized with a two-layer GCN, making the BDR-CNN-GCN model employed to analyze breast mammograms using a 14-way data augmentation approach. The results showed that the BDR-CNN-GCN model achieved an accuracy of 96% when analyzing breast mammograms using the MIAS database.

The approaches proposed in the related work present methods and models to improve the accuracy of description and classification. In our work, we have concentrated on describing the spiculated masses to ameliorate the accuracy of CADx systems.

3. PATAR DESCRIPTOR

As mentioned previously, the main goal of our work is to provide radiologists with a robust descriptor that helps them in the diagnosis procedure. An efficient CADx system is based on a solid descriptor with significant features. Due to the random shape of masses and the high resemblance between breast tissue and masses, the description of spiculated masses remains a difficult task and an unsolved problem in the domain of breast cancer. In the literature, most of the descriptors developed until now cannot extract all spiculated parts of masses and automatically false the classification step [22]. To overcome this problem, we proposed an approach based on two steps. First, we begin by applying a geometric transformation, a polygon approximation of the mass contour. The polygon approximation isolates and well-estimates concave and convex areas, which are the key properties that distinguish benign from malignant masses. Secondly, to calculate the size of spiculations of masses, Triangle-Area Representation (TAR) is used. All the corners made by the polygon approximation are browsed in the clockwise direction to calculate the triangles formed by three consecutive points. Our descriptor is based on polygon approximation as the first step, followed by the TAR calculation, making a robust descriptor capable of detecting spiculation in masses and calculating their size. Figure 2 explains the functioning of the PATAR descriptor; each step is detailed in the sections below.

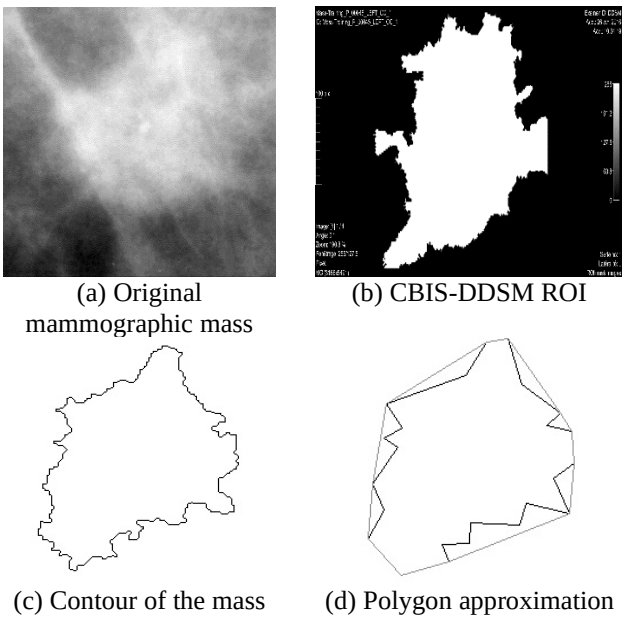
Figure 2. Illustration of PATAR descriptor, polygon approximation is applied on mass ROI's then TAR signature to extract features

3.1 Polygon approximation

Polygonal approximation simplifies complex shapes or contours by representing them using polygons with minimum

vertices. The goal is to maintain important features of the original form while reducing the level of detail. This simplification is often done for computational efficiency, data compression, or visualization [23]. Polygonal approximation enhances the efficiency of shape characterization and classification. This method is efficient because of its strong representation and insensitivity to translation, scale, and rotation invariance. These significant features are found helpful in many applications. The primary goal of polygon approximation is to represent a curve using a polygonal shape, with its vertices determined by a selected subset of points along the curve.

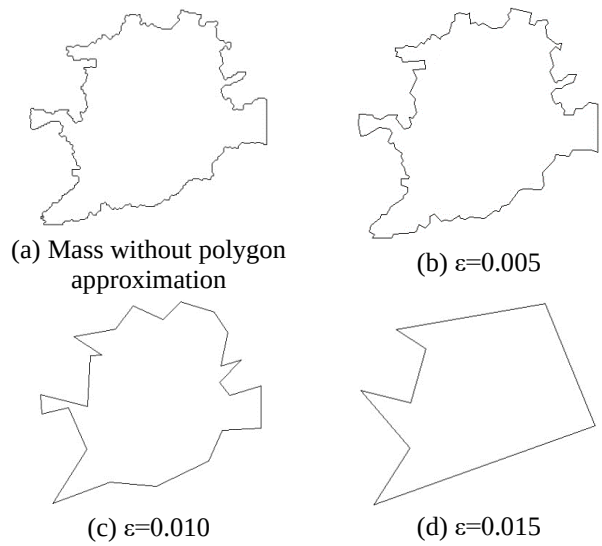
In our context, polygon approximation plays a crucial role in maintaining the degree of spiculation of masses while reducing contours to a polygon form. Figure 3 shows an example of polygon approximation applied to digital mammographic mass.



(a) Original mammographic mass (b) CBIS-DDSM ROI

(c) Contour of the mass (d) Polygon approximation

Figure 3. Example of Polygon approximation: a) the original mammographic mass, b) CBIS-DDSM ROI, c) contour of the mass and d) the result of its polygon approximation



(a) Mass without polygon approximation (b) ε=0.005

(c) ε=0.010 (d) ε=0.015

Figure 4. Different values of parameter ε on same mass contour, (a) mass without polygon approximation (b) ε=0.005 (c) ε=0.010 and (d) ε=0.015

Algorithm 1. Douglas-Peucker polygon approximation algorithm

```
procedure DouglasPeucker(PointList[1...n], tolerance: real)
  dmax := 0
  index := 0
  for i := 2 to n - 1 do
    d := PerpendicularDistance(PointList[i],
Line(PointList[1], PointList[n]))
    if d > dmax then
      index := i
      dmax := d
  end for
  if dmax > tolerance then
    recResults1 := DouglasPeucker(PointList[1...index],
tolerance)
    recResults2 := DouglasPeucker(PointList[index...n],
tolerance)
    return concatenate(recResults1, recResults2)
  else
    return Line(PointList[1], PointList[n])
  end if
end procedure
```

The Ramer-Douglas-Peucker algorithm [23] is used for the polygon approximation method. Let us consider the curve Cd = {p1, p2, …., pn}, where pi=(xi, yi) are points in the clockwise direction in the discrete 2-dimensional space. These curves are derived from the edges of mammographic masses using contour-detection techniques. The algorithm starts by identifying the start and end points of the mass contour. In their paper, Douglas and Peucker [23] refer to these two points as the anchor point and the floating point. The algorithm has one parameter, ε, and the value of ε defines the degree of approximation. In Figure 4, different values of ε are used for the same mass, showing the importance of ε in keeping the most significant spiculations of masses. More ε parameter is important, more variations in contour are ignored, and the polygon is less representing the original form of mass. In Algorithm 1, tolerance refers to the ε parameter.

The value of the ε parameter is fundamental for the PATAR descriptor; a false value may omit the concave and convex spaces in mass. The challenge is to find a value of ε that keeps the spiculated part of the contour's mass without deformity and, on the other hand, transforms a mass into a polygon that preserves information and characteristics of masses. In this work, ε parameter of the polygon approximation function is fixed to 0.01 using the Douglas-Peucker algorithm [23]. ε=0,01 is the universal value and guaranty the best representation of the mass by preserving the morphology of the shape.

3.2 TAR signature

Concavity and convexity are shape features made by spiculations in masses; they are the most significant characteristics used to discriminate between malignant or benign masses. This approach uses triangle-area representation (TAR) calculation in the second step of the PATAR descriptor after polygon approximation. In this paper, a triangle-area representation was used to detect and calculate concavity and convexity with high precision. The TAR function calculates the area formed by triangles shaped by the mass contour. The curvature of the contour point (i_n, j_n) is

calculated using the TAR function, defined as follows:

$$TAR\ (n, ts) = \frac{1}{2}\begin{vmatrix} i_{n-ts} & j_{n-ts} & 1 \\ i_n & j_n & 1 \\ i_{n+ts} & j_{n+ts} & 1 \end{vmatrix} \quad (1)$$

For every three consecutive pixels of the contour of masses $P_{n\text{-}ts}$ ($i_{n\text{-}ts}$, $j_{n\text{-}ts}$), P_n (x_n, y_n), and P_{n+ts} (x_{n+ts}, y_{n+ts}), where $n \in [1, N]$ and $t_s \in [1, N/2 - 1]$, t_s represent the step of the TAR function if t_s=1 means that P_{n+1} is the neighbor of P_n in the clockwise direction and P_{n-1} is the neighbor in counter-clockwise direction. The triangle is formed by the points P_{n-1}, P_n, and $P_{n+1}$is given by Alajlan et al. [24]. A pseudo-code of the TAR signature is presented in Algorithm 2.

Algorithm 2. TAR signature of contour

```
Procedure TAR(PointList[1...n], step: integer)
for p := 1 to n do
        TAR[p] := det[(p_{i-step},p_{j-step} ,1),(p_i,p_j,1),(p_{i+1},p_{j+1},1)]
if TAR(p) < 0 then p is concave point
if TAR(p) > 0 then p is convex point
if TAR = 0 traight line
return TAR[]
end procedure
```

Figure 5 illustrates that the TAR function can generate three distinct outcomes: negative, zero, and positive values when the contour is traversed in a counter-clockwise direction. These outcomes signify the nature of the area enclosed by the three points, indicating whether it is concave, straight-line, or convex [24]. Additionally, the TAR signature exhibits significant invariance, meaning it remains consistent even when subjected to translation, rotation, and scaling operations. This dual advantage of efficiency and invariance makes triangle-area representation a valuable tool in our research, where it aids in rendering smooth curves accurately while reducing computational overhead.

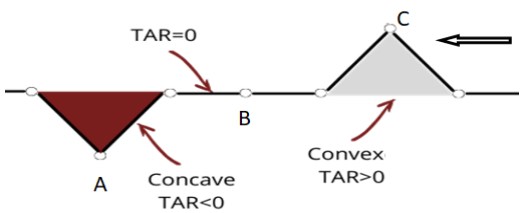


Figure 5. Three possible values of TAR signature, convex when TAR is positive, convex when TAR is negative and straight line if TAR=0

3.3 Feature extraction

The role of a descriptor in computer vision and image processing is to capture and represent critical characteristics or features of an image quantitatively and mathematically. Descriptors try to simulate human perception by quantifying visual features and patterns in images, aiming to replicate how humans interpret and understand visual content. A *feature* is a measurable information extracted from an image to identify and classify objects. PATAR descriptor aims to quantify spiculation in mass and qualify its degree to assign mass in a specific class (Benign or malignant). PATAR takes the ROI image in input and provides a series of features generated by the implemented approach.

The malignant mass has a specific topology characterized by lines of varying length and thickness radiating from the margin of the mass. As mentioned before, PATAR aims to isolate concave and convex spaces and begin the process by applying a polygon approximation. All the corners in the polygon generated are marked to form the triangles and calculate their sizes with triangle-area representation (TAR). The corners are browsed in the clockwise direction; for each point, P_i, the area of the triangle formed by P_i, P_{i-1}, and P_{i+1} is measured. The number of corners, the concave, and convex, are the extracted features using the PATAR descriptor.

- **Number of corners:** a shape with high irregularity contains more corners compared to round shapes. Oval and round masses are usually benign and contain fewer corners than irregular shapes. Figure 6 and Figure 7 illustrate the relation between the number of corners and the probability of roundness of mass shape.

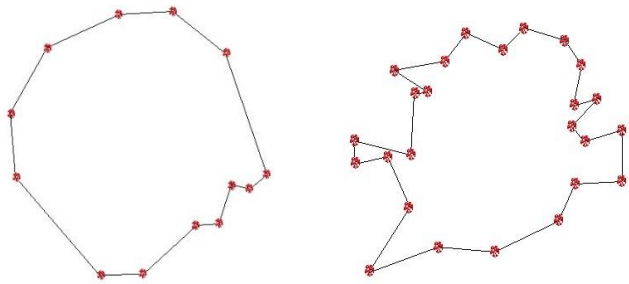


Figure 6. (a) Benign mass with round shape having 13 corners, (b) malignant mass with irregular shape having 25 corners

- **Negative and Positive TAR:** TAR signature is used to extract and evaluate convex and concave spaces formed by the polygon's corners. The triangles are separated according to their values, as follow:

$$\begin{cases} TARN = \sum_{i=1}^{n} TARP_i < 0 \\ TARP = \sum_{i=1}^{n} TARP_i > 0 \end{cases} \quad (2)$$

n is the number of corners, the area of triangles i.e., the values of *TARN* and *TARP* made by the corners of shapes define the degree of spiculation in mass. A high value of *TARN* and *TARP* means the mass is probably malignant. Round mass is characterized by a value of *TARN* close to zero.

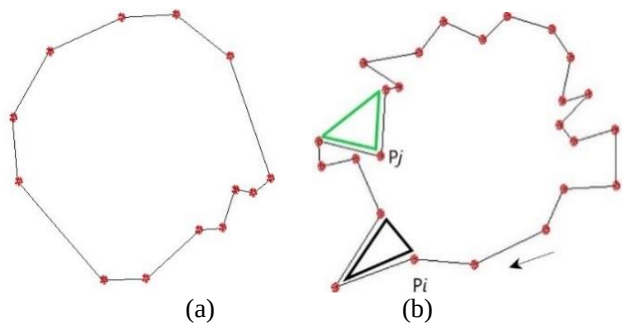


Figure 7. Mass (a) and (b) shows the difference between round (a) and spiculated mass (b) in term of *TARN* (Green triangle) and *TARP* (black triangle) values

In addition to TARN, TARP and the number of corners,

mass area, and contour length are used as features in the classification process. The extracted features are formulated in the following vector PATAR [TARN, TARP, NBP, MA, CL].

4. MASS CLASSIFICATION

The features obtained during the description phase serve as inputs for the subsequent classification process. The classification aims to allocate each mass to the suitable class, benign or malignant. In the domain of breast cancer, every decision made by a radiologist carries a risk of error, and only a biopsy can decide definitively if a mass is benign or malignant. For those reasons we choose a fuzzy classifier in this stage. A fuzzy classifier will attribute for each mass a probability of membership to every class. In addition to Fuzzy C-means, SVM and Random Forest are selected to perform classification.

This step uses three classifiers to evaluate the PATAR descriptor: Fuzzy C-means (FCM), SVM, and Random Forest (RF). The choice of classifiers depends on the data types, scales, and classification purpose.

Fuzzy C-means (FCM) is a soft clustering method that assigns each data point in a dataset to N clusters with a certain degree of membership or probability score, indicating the likelihood of the data point belonging to each cluster [25]. In the literature, fuzzy classification is not widely used in computer-aided diagnosis systems (CADx). A fuzzy classifier in breast cancer diagnosis shows realistic results. In reality, a radiologist cannot define if a mass is malignant or benign, and only a biopsy gives a 100% diagnosis; for these reasons, a probability of malignancy and benignity should be attributed to each mass. In our approach, the probability of membership is equal to the distance of each point (Features vector of the image) to the centroid of the class (Malignant or benign). Based on the ground truth training dataset, N masses are divided into two classes, malignant and benign J_a and J_b, respectively, and the matrix membership U_{ia}, $U_{ib} \in \{0,1\}$ i=1, …, N is also created. For each class, the centroid is calculated as follows [25]:

$$C_{a,b} = \frac{\sum_{i=1}^{N} U_{i(a,b)}^{m} x_i}{\sum_{i=1}^{N} U_{i(a,b)}^{m}} \quad (3)$$

In this classification stage, the distance between centroid classes (Benign and malignant) and the mass point is calculated for every mass in the test dataset. Based on those distance membership to each class is estimated. The membership $U_{i(a,b)}$ of a mass defines their probability of malignity and benignity. $U_{i(a,b)}$ is calculated as follows [25]:

$$U_{i(a,b)} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_{a,b}\|}{\|x_i - c_{a,b}\|} \right)^{\frac{2}{m-1}}} \quad (4)$$

Support vector machines as the supervised classifier are widely used, especially for two-class problems like mass classification. SVM provides distance to the hyperplane instead of probability as FCM and random forest. Random Forest is a very powerful classifier when features are on various scales.

5. EXPERIMENTS AND RESULTS

The adopted strategy in experiments is separated into two phases: learning and testing. In the training part, features are extracted from the ROIs of the CBIS-DDSM dataset. ROIs are a portion of the images containing the abnormality; these are delimitated and annotated by mammographs and radiologists in the CBIS-DDSM dataset. CBIS-DDSM is an improved and standardized version of DDSM designed for evaluating CAD systems [26]. Using the ground truth of the CBIS-DDSM dataset, we evaluate the PATAR descriptor. Figure 8 shows the outline of our CADx system based on PATAR.

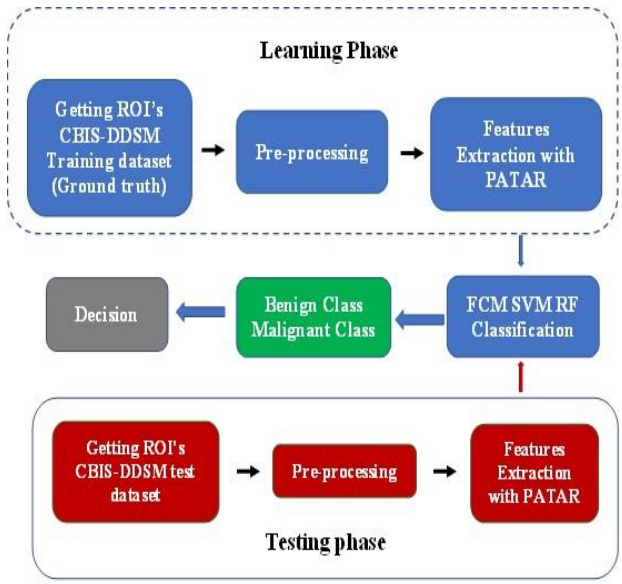


Figure 8. Overview of the proposed approach, in the learning phase training dataset is used to build the feature's model. SVM, FCM and RF classifiers are used to distinguish malignant from benign masses. Finally, the testing to evaluate PATAR

Our approach for automatic diagnosis was put to the test through a series of experiments. These experiments evaluated its performance by measuring accuracy, sensitivity, specificity, and the F1-score. Python software (version 3.7) with OpenCV library was used on a PC with Intel i5 (2.00 Ghz) with 8 GB of RAM and a Windows 10 operating system. CBIS-DDSM was used to assess the model's performance [26]. The dataset is separated into two subsets, one for the training containing 1318 ROIs (637 Malignant and 681 benign masses) and another for the test and validation step composed of 378 ROIs (147 Malignant and 231 benign masses).

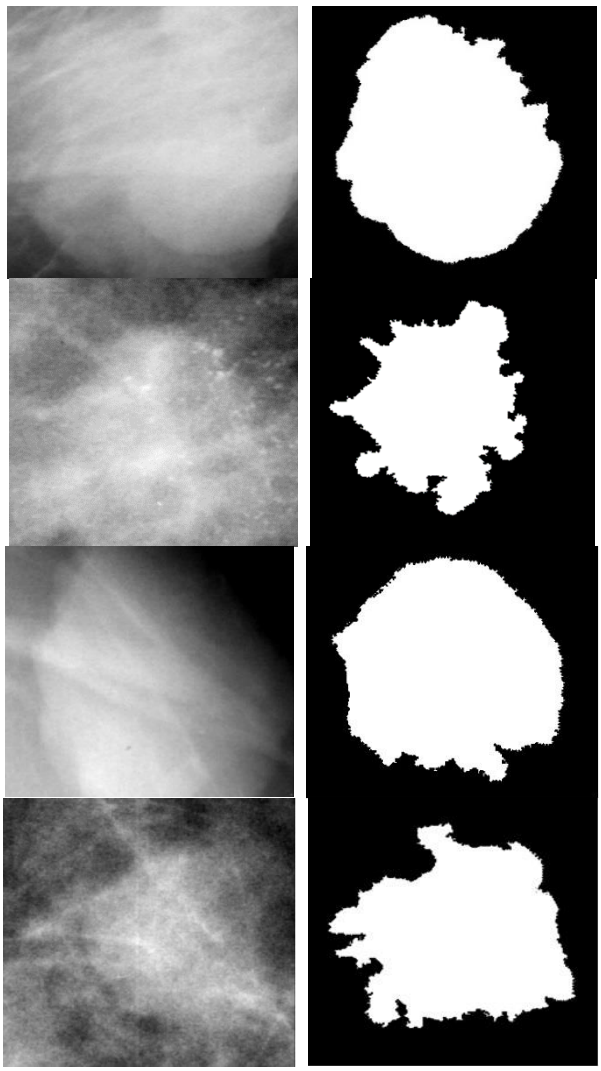
5.1 CBIS-DDSM dataset

The Digital Database for Screening Mammography (DDSM) contains digitized images from scanned mammography films compressed with lossless JPEG encoding. In these experiments, we have used a database version, i.e., CBIS-DDSM, containing an updated and standardized version subset of the original DDSM images in the standard Digital imaging and communications in medicine (DICOM) format. Table 1 shows the number of cases, abnormalities, and images in each set, training, and testing. Masses are located and approved by an experiment radiologist [26].

Table 1. Number of Cases, Abnormalities and Images in the Training and Test Sets, each case can have one or more abnormalities and more images

	Benign Cases			Malignant Cases			
	Cases	**Abnormalities**	**Images**	**Cases**	**Abnormalities**	**Images**	**Total Images**
Training	355	387	681	336	361	637	1318
Testing	117	135	231	83	87	147	378

Computer-aided diagnosis systems (CADx) need only analyze regions of interest (ROIs), not full mammogram images. ROIs feature abnormalities within the cropped sections of the image, which outline the bounding rectangle of the abnormality relative to its ROI see Figure 9. Our descriptor performed calculations directly on masses segmented and delineated the mass from the enveloping tissue. Ground truth provided by CBIS-DDSM is founded on the BI-RADS category.



(a) (b)

Figure 9. Four cropped images of mass from CBIS-DDSM dataset, a) mammographic images without segmentation, b) shows segmented mass (Mask image)

5.2 Results and discussion

In order to determine the effectiveness of the proposed approach, 1 545 mammograms were used, 1 318 (681 benign and 637 malignant) images were in the training phase, and 247 images were selected for testing (131 benign and 96 malignant). Our method's diagnostic performance is assessed regarding computing time, sensitivity, specificity, and accuracy. Accuracy measures how many test cases the classifier correctly classifies, while *sensitivity* (SN) represents the valid positive rate, and *specificity* (SP) denotes the true negative rate. These parameters are formally defined as follows:

$$SN = \frac{TP}{TP+FN} \tag{5}$$

$$SP = \frac{TN}{TN+FP} \tag{6}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{7}$$

where, TP is the true positive, FP is the false positive, TN is the true negative, and FN is the false negative.

The obtained accuracy in this method proves that combining polygon approximation and TAR signature improves the classification of mammograms in order to diagnose breast cancer on digital mammograms automatically. Polygon approximation is valuable in our descriptor, especially in simplifying complex masse shapes or contours in mammographic images. Breast mass can contain intricate shapes and contours, especially spiculated masses. Polygon approximation simplifies these shapes by representing them with reduced vertices without losing their characteristics (Spiculations) and making triangle-area representation (TAR) more beneficial. Without polygon approximation, TAR calculations are done on all negligible contour variations and false completely the description and poorly estimate the spiculations.

In Table 2, the results of the comparison between our approach and classification without polygon approximation using TAR signature. The accuracy obtained with polygon approximation and Fuzzy C-means classifier is 96.76%, 97,94% with Random Forest, and 94,65% with Support vector machines. Without polygon approximation, the classification accuracy decreases significantly to 82.80%. The amelioration gained in terms of accuracy, precision, sensitivity, and specificity with polygon approximation confirms with exactitude our hypotheses at the beginning of our paper that spiculation in masses will be well raised, represented, extracted, and evaluated with our contribution using PATAR descriptor. Compared results are presented in Figure 10.

The classification performance using the DDSM dataset for the three classifiers (RF, FCM, and SVM) is indicated in the ROC plots in Figure 11. The area under a receiver operating characteristic (ROC) curve, abbreviated as AUC, offers an aggregate performance measure across various classification thresholds. The AUC (Area Under the Curve) is an evaluation metric ranging from 0 to 1, with 1 indicating the highest level of performance. The SVM classifier's AUC of FCM was 95,48 and 96,23; Random Forest gives the best performance with 97,13 of AUC and 97,94 accuracy.

Table 3 compares our approach and some existing works on

mammographic classification to better demonstrate the performance of our method. Our work gives one of the best results in accuracy, sensitivity, and specificity using a maximum number of mammograms.97.94% is the highest accuracy obtained with a random forest classifier using more than double the samples (mammograms) in most of the studies presented in Table 3.

Table 2. Comparison of results obtained with and without polygon approximation

	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
With Polygon Approximation					
RF	97.94	98.31	93.55	99.45	95.87
FCM	96,76	95.08	92.06	98.37	93.55
SVM	94,65	98,04	80,65	99,45	88,50
Without Polygon Approximation	82.80	76.29	77.32	83.57	74.81

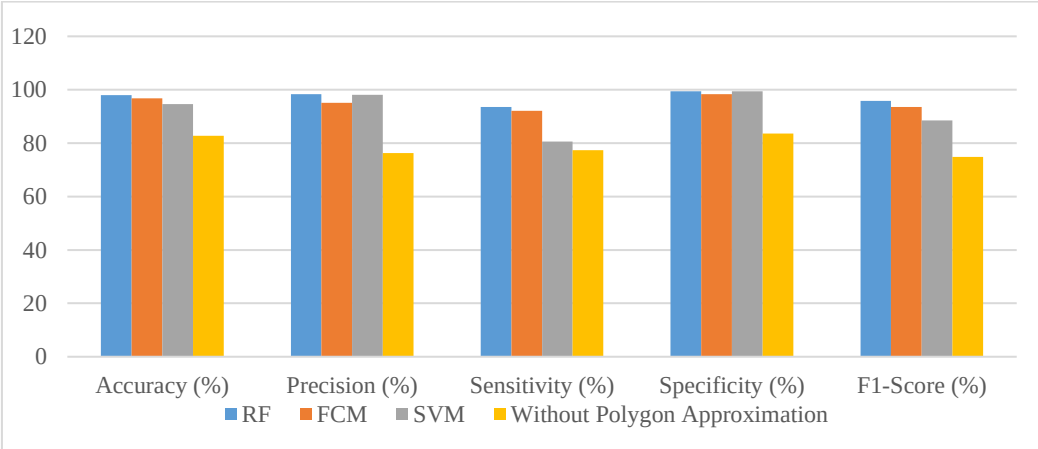
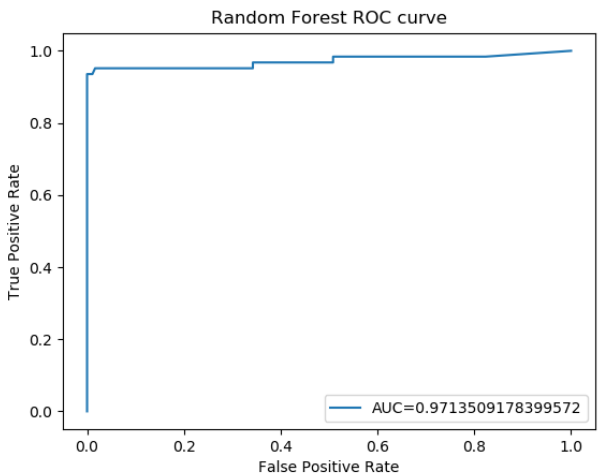
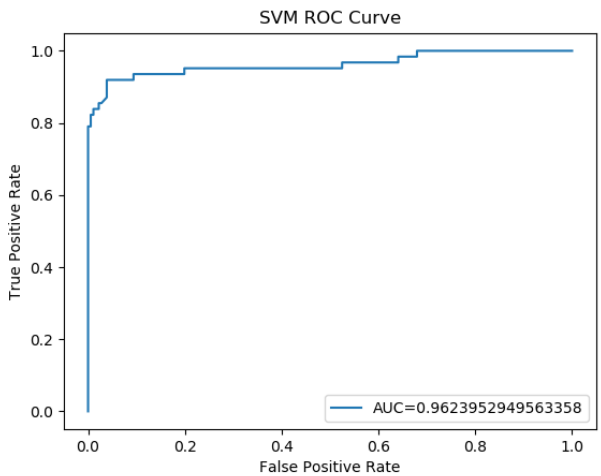


Figure 10. Comparing the best results of PATAR, with and without polygon approximation

Table 3. Comparison between proposed approach and some previous works

METHOD	Number of Mammograms	Database	Accuracy (%)	Sensitivity (%)	Specificity (%)	Classifier	Year
PATAR – FCM	1545	CBIS-DDSM	96.76	92.06	98.37	FCM	-
PATAR – SVM	1545	CBIS-DDSM	94.65	80.65	99.45	SVM	
PATAR – RF	1545	CBIS-DDSM	97.94	93.55	99.45	RF	
Arora et al. [20]	-	CBIS-DDSM	88	-	-	SVM	2020
Pezeshki et al. [18]	200	DDSM	93.22	92.06	94.54	SVM	2019
Vijayarajeswari et al. [17]	322	MIAS	94	-	-	SVM	2019
Goudarzi et al. [15]	-	mini-MIAS	89.37	88.23	84.23	Fuzzy	2018
Souza et al. [12]	620	DDSM	92.15	91.40	92.90	SVM	2017
Xie et al. [11]	300	DDSM	96.02	94.88	97.16	SVM	2016
Beheshti et al. [10]	168	DDSM	87.81	97.37	79.55	SVM	2016





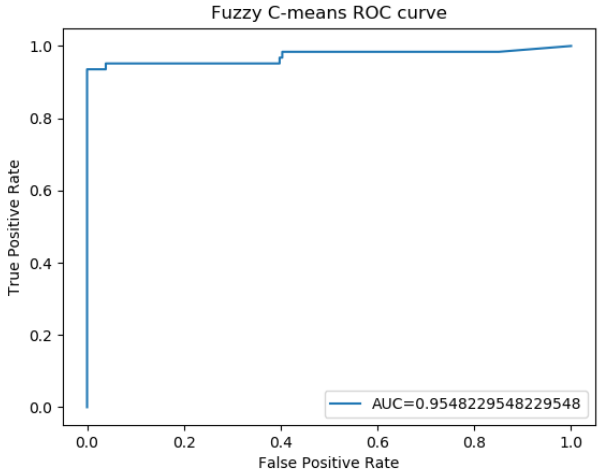


Figure 11. ROC curve of RF, SVM and FCM using CBIS-DDSM. SVM classifier ensures the best results in terms of accuracy and AUC

6. CONCLUSION

In this paper, a new mass descriptor is proposed for breast cancer diagnosis using digitized mammograms. PATAR descriptor performs a polygon approximation on masses to eliminate slight variation in contours and detect important spiculations, which are significant characteristics for automatic diagnosis in the breast cancer domain. Finally, Fuzzy C-means, SVM, and random forest were utilized for classification. The results obtained on mammograms from CBIS-DDSM confirm the utility of our contribution. With an accuracy of 97.94%, the PATAR descriptor presents one of the best results using a random forest classifier. We recommend conducting experiments using different image databases to improve the method's validation. The epsilon (ε) parameter can be revisited by applying a generic method that calculates the optimal value of (ε).

REFERENCES

[1] World Health Organization International, Agency of Research on Caner: PRESS RELEASE N° 292, 15 December 2020.

[2] Singh, L., Jaffery, Z.A. (2018). Computerized detection of breast cancer in digital mammograms. International Journal of Computers and Applications, 40(2): 98-109. https://doi.org/10.1080/1206212X.2017.1395131

[3] Nass, S., Henderson, I., Lashof, J. (2001). Committee on technologies for the early detection of breast cancer. In Mammography and Beyond: Developing Technologies for the Early Detection of Breast Cancer. Nat. Cancer Policy Board, Inst. Med., Commission Life Stud, Nat. Res. Council.

[4] Burnside, E.S., Sickles, E.A., Bassett, L.W., Rubin, D.L., Lee, C.H., Ikeda, D.M., Mendelson, E.B., Wilcox, P.A., Butler, P.F., D'Orsi, C.J. (2009). The ACR BI-RADS® experience: Learning from history. Journal of the American College of Radiology, 6(12): 851-860. https://doi.org/10.1016/j.jacr.2009.07.023

[5] Berment, H., Becette, V., Mohallem, M., Ferreira, F., Chérel, P. (2014). Masses in mammography: What are the underlying anatomopathological lesions? Diagnostic and Interventional Imaging, 95(2): 124-133. https://doi.org/10.1016/j.diii.2013.12.010

[6] Madhukar, B.N., Bharathi, S.H., Polnaya, A.M. (2023). Multi-scale convolution based breast cancer image segmentation with attention mechanism in conjunction with war search optimization. International Journal of Computers and Applications, 45(5): 353-366. https://doi.org/10.1080/1206212X.2023.2212945

[7] Kaushal, C., Bhat, S., Koundal, D., Singla, A. (2019). Recent trends in computer assisted diagnosis (CAD) system for breast cancer diagnosis using histopathological images. IRBM, 40(4): 211-227. https://doi.org/10.1016/j.irbm.2019.06.001

[8] Katzen, J., Dodelzon, K. (2018). A review of computer aided detection in mammography. Clinical Imaging, 52: 305-309. https://doi.org/10.1016/j.clinimag.2018.08.014

[9] Guliato, D., Rangayyan, R.M., de Carvalho, J.D., Santiago, S.A. (2006). Spiculation-preserving polygonal modeling of contours of breast tumors. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, pp. 2791-2794. https://doi.org/10.1109/IEMBS.2006.260441

[10] Xie, W., Li, Y., Ma, Y. (2016). Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing, 173: 930-941. https://doi.org/10.1016/j.neucom.2015.08.048

[11] Beheshti, S.M.A., Noubari, H.A., Fatemizadeh, E., Khalili, M. (2016). Classification of abnormalities in mammograms by new asymmetric fractal features. Biocybernetics and Biomedical Engineering, 36(1): 56-65. https://doi.org/10.1016/j.bbe.2015.07.002

[12] Souza, J.C., Silva, T.F.B., Rocha, S.V., Paiva, A.C., Braz, G., Almeida, J.D.S., Silva, A.C. (2017). Classification of malignant and benign tissues in mammography using dental shape descriptors and shape distribution. 7th Latin American Conference on Networked and Electronic Media (LACNEM 2017), pp. 22-27. https://doi.org/10.1049/ic.2017.0030

[13] Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D. (2002). Shape distributions. ACM Transactions on Graphics (TOG), 21(4): 807-832. https://doi.org/10.1145/571647.571648

[14] Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I. (2018). Detecting and classifying lesions in mammograms with deep learning. Scientific Reports, 8(1): 4165. https://doi.org/10.1038/s41598-018-22437-z

[15] Goudarzi, M., Maghooli, K. (2018). Extraction of fuzzy rules at different concept levels related to image features of mammography for diagnosis of breast cancer. Biocybernetics and Biomedical Engineering, 38(4): 1004-1014. https://doi.org/10.1016/j.bbe.2018.09.002

[16] Sun, L., Wang, J., Hu, Z., Xu, Y., Cui, Z. (2019). Multi-view convolutional neural networks for mammographic image classification. IEEE Access, 7: 126273-126282. https://doi.org/10.1109/ACCESS.2019.2939167

[17] Vijayarajeswari, R., Parthasarathy, P., Vivekanandan, S., Basha, A.A. (2019). Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Measurement, 146: 800-805. https://doi.org/10.1016/j.measurement.2019.05.083

[18] Pezeshki, H., Rastgarpour, M., Sharifi, A., Yazdani, S.

(2019). Extraction of spiculated parts of mammogram tumors to improve accuracy of classification. Multimedia Tools and Applications, 78: 19979-20003. https://doi.org/10.1007/s11042-019-7185-4

[19] de Brito Silva, T.F., de Paiva, A.C., Silva, A.C., Braz Júnior, G., de Almeida, J.D.S. (2020). Classification of breast masses in mammograms using geometric and topological feature maps and shape distribution. Research on Biomedical Engineering, 36: 225-235. https://doi.org/10.1007/s42600-020-00063-x

[20] Arora, R., Rai, P.K., Raman, B. (2020). Deep feature–based automatic classification of mammograms. Medical & Biological Engineering & Computing, 58: 1199-1211. https://doi.org/10.1007/s11517-020-02150-8

[21] Zhang, Y.D., Satapathy, S.C., Guttery, D.S., Górriz, J.M., Wang, S.H. (2021). Improved breast cancer classification through combining graph convolutional network and convolutional neural network. Information Processing & Management, 58(2): 102439. https://doi.org/10.1016/j.ipm.2020.102439

[22] Kohli, A., Jha, S. (2018). Why CAD failed in mammography. Journal of the American College of Radiology, 15(3 Pt B): 535-537. https://doi.org/10.1016/j.jacr.2017.12.029

[23] Douglas, D.H., Peucker, T.K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartographica: The International Journal for Geographic Information and Geovisualization, 10(2): 112-122. https://doi.org/10.3138/FM57-6770-U75U-7727

[24] Alajlan, N., El Rube, I., Kamel, M.S., Freeman, G. (2007). Shape retrieval using triangle-area representation and dynamic space warping. Pattern Recognition, 40(7): 1911-1920. https://doi.org/10.1016/j.patcog.2006.12.005

[25] Bezdek, J.C., Ehrlich, R., Full, W. (1984). FCM: The fuzzy C-means clustering algorithm. Computers & Geosciences, 10(2-3): 191-203. https://doi.org/10.1016/0098-3004(84)90020-7

[26] Curated Breast Imaging Subset of Digital Database for Screening Mammography. http://doi.org/10.7937/K9/TCIA.2016.7O02S9CY