Automated Feature Extraction from Breast Masses using Multiscale Fractal Dimension

José Robson de Souza Filho, Carolina Yukari Veludo Watanabe
Computer Science Department
Federal University of Rondônia - UNIR
Porto Velho, RO, Brazil
jrobsonsf@gmail.com, carolina@unir.br

***Abstract*—This paper proposes a new computer-aided diagnosis method to characterize benign and malignant masses in mammograms. First, for each image, a region of interest is segmented using an improved version of the EM/MPM algorithm. Then the contour is obtained by applying the Sobel high-pass filter. To extract the features, we compute the fractal dimension of the contour using the Bouligand-Minkoswki technique, with several successive dilations. This results in a curve, to which is applied multiscale differentiation, and twenty-two inflection points are obtained. The Y-axis coordinates are used to compose the feature vector. To select features, an attribute worth evaluator based on the 1R classifier is applied; it ranks three main inflection points that compose the final descriptor. This descriptor is submitted to the LADtree classifier, which finally suggests the diagnosis. The result of comparing the proposed method with traditional descriptors shows that it is well-suited to characterizing mammograms.**

Keywords—feature extraction; breast cancer; multiscale fractal signature; LADtree

I. INTRODUCTION

Breast cancer is the second most common type of cancer in the world and the most common among women. It is responsible for about 28% of new cancer cases per year and is the main cause of cancer death in the Brazilian female population. In 2016, the National Institute of Cancer (INCA) estimated 57,960 new cases of breast cancer in Brazil [1]. Fortunately, this deadly condition has a high rate of curability as long as it is diagnosed in the early stages. But it is still a challenging task for the radiologists to distinguish between malign and benign mass. Abnormal cases have varied contour shapes, textures, and sizes. It is very difficult even for experienced radiologists to discriminate whether the breast mass is malign [2]. Currently, mammography is still the most widely used and accurate diagnostic tool for breast cancer, and computer-aided detection techniques have been developed to support radiologists in the their diagnostic work [3].

There are two main challenges in Computer-Aided Diagnosis (CAD) systems. The first one is to find appropriate features to map low-level visual features to high-level semantic concepts (*semantic gap*) [2]. The second one is the *curse of high dimensionality* [4], in which a high increase in the number of features (and consequently in the dimensionality of the data) leads to losing the significance of each feature value. Thus, to avoid reducing the accuracy of the discrimination, it is preferred to keep the number of features as low as possible, establishing a trade-off between discriminatory power and the feature vector size.

A typical benign mass has a contour that is round, smooth, and well-defined, whereas a typical malignant tumor has a contour that is spiculated, rough, and ill-defined. The significant differences between the boundary shape characteristics of benign masses and malignant tumors may be used to differentiate them by deriving shape factors [5]. Therefore, the fractal dimension may be used to quantify the complexity or irregularity of an object's boundary.

In this paper, we present a computer-aided diagnosis method that combines a multiscale fractal approach to characterize the shape of the tumors, with a feature selection algorithm, aimed at surmounting the challenges described above and keeping in mind the power of shape representation of the fractal approach. The method is divided into two steps: training and testing. In the training step, each image of the training dataset entails the application of an automatic method to segment the image and select the region of interest (tumor) using an algorithm based on the EM/MPM (Expectation-Maximization/Maximization of the Posterior Marginals) technique. Next, the contour is obtained and a multiscale fractal approach is adopted, based on successive dilations of the contour, and the fractal dimension is computed for each dilation. These values form a curve. From this curve several inflection points are obtained, the Y-coordinates being considered to compose the feature vector. The feature vectors of all the images from the training dataset are submitted to a feature selection algorithm based on the 1R (One Rule or OneR) classifier [6], and the three main points to compose the final descriptor are chosen. These descriptors are used to build a decision tree using the LADtree (LogitBoost Alternating Decision Tree) classifier [7], which is responsible for classifying new cases. In the test step, suggestions of diagnosis of new cases may arise. An image is submitted to the processing phase to generate its feature vector, which is submitted to the LADtree classifier.

To evaluate our proposal, we performed experiments comparing the results with well-known descriptors in the literature (Zernike moments [8] and a feature extractor based on shape analysis [9]).

0-7695-6357-0/17/$31.00 ©2017 IEEE
DOI 10.1109/WVC.2017.00012

Our method potentially generates infinite points of interest, which can be analyzed to find the most relevant ones and whose coordinates are used in the formation feature vectors. The results of our research revealed that the overall accuracy of the classification resulting from this technique is directly related to the selected points of interest, and our method yielded high precision in the correct identification of malignant tumors, and the best values of accuracy selecting only three features.

The remainder of this paper is organized as follows. Section II introduces the basic concepts for developing the work. In Section III concerns the proposed method. Section IV presents the experiments and results for a mammogram image dataset, and in the Section V some conclusions are drawn.

II. Background and Related Works

In recent years, several tumor classification techniques have been established to assist radiologists in diagnosing breast cancer. Among the most often used, highly accurate, descriptors for mass characterization in the literature are Zernike moments and their variations [8, 10-12].

Another technique that has been explored is the fractal method. The idea that objects have fractional dimensions has been discussed since the 19th century [13]. The main idea is that in fractal geometry objects have an intermediate dimension and not a natural number, as straight lines, planes and solids, referring to 1-, 2- and 3- dimensional figures, respectively. In the image analysis paradigm, fractal dimension measurements are used to estimate and quantify the complexity of form, shape and texture of objects [13].

Fractal analysis has been shown to be useful in image processing for characterizing shape and gray-scale complexity. Mammograms [14] of breast masses present shapes and gray-scale characteristics that vary between benign masses and malignant tumors. However, few studies have been conducted on the application of fractal analysis specifically for classifying breast masses based on shape [5].

We found some fractal methods which have been used to shape analysis, based on the contour of a region of interest to differentiate between benign and malignant cases in mammograms, as follows.

Dobrescu et al. [15] reported on a morphological study of 30 cases computing two measures: the fractal dimension, using a box-counting algorithm, and the lacunarity for each mammogram texture. They achieved high accuracy taxes using n threshold to classify the cases. However, this threshold is chosen by observing the dataset distribution when using these measures; the method is not completely automatic, since the contour of each mass is manually drawn by an expert radiologist specialized in mammography; and their experiments were made on a small number of breast cancer cases.

Beheshti et al. [16] proposed an image segmentation fractal method to detect masses in mammograms, defining new fractal dimensions to calculate the threshold for discriminating mass from background tissue. From the contour, they computed three features to compose the feature vector: N2, SMF e A1, which are also based on the fractal theory. In the task of classification, they used the SVM algorithm.

Li et al. [2] proposed a method to express the regularity of the contour for breast mass. First, the region of interest was labelled manually by expert doctors. Second, they translated a 2D contour to 1D signature using the Euclidean distance from the edge of the breast mass to the periphery of the circular or oval center with centroid, and the whole 1D signature into different subsections was segmented. From each subsection, four local features were extracted: root mean square (RMS) roughness which describes the irregular degree of 1D signature; the μ_R/σ_R ratio which describes the circularity of the breast mass contour; the fractal dimension computed by the Brownian motion model; and the RMS slope. 1-Nearest-Neighbor (1-NN), Support Vector Machine (SVM) and Artificial Neural Network (ANN) were used as classifiers.

All these studies emphasize the usefulness of fractal parameters in tumor pathology. Although various algorithms exist for computing fractal dimensions for digital imaging, our approach is based on the multiscale Bouligand-Minkowski [13] method.

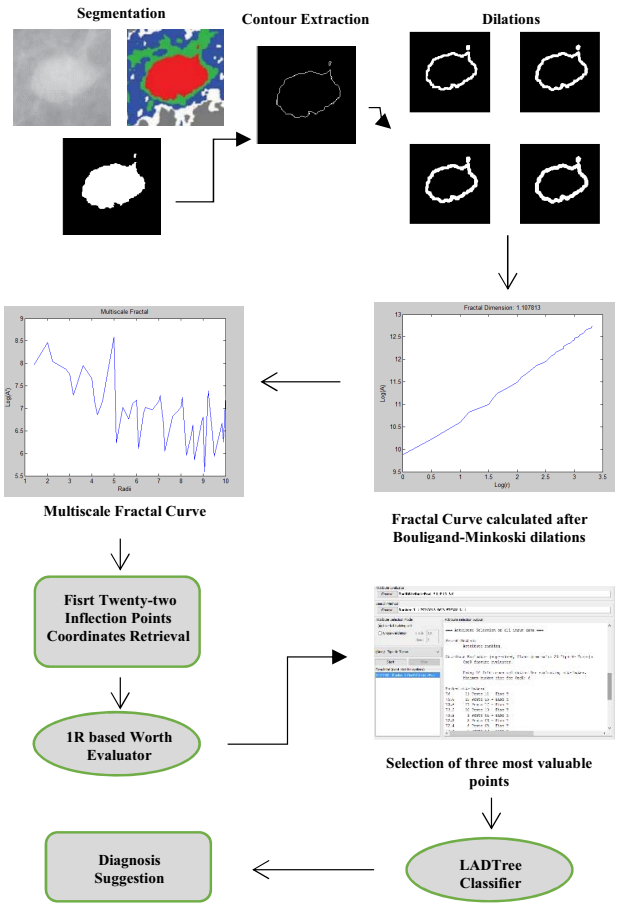


Fig. 1. Pipeline of the proposed method

III. METHODOLOGY

In this paper, we propose a shape descriptor based on the multiscale Bouligand-Minkowski technique [17] and the 1R algorithm. The pipelines of the method are presented in Fig. 1.

First, a dataset was collected, containing benign and malign masses. Then an improved version of EM/MPM algorithm was applied to each image to detect the regions of interest (masses). Next the contour of each was extracted. A multiscale approach based on the Bouligand-Minkowski method adopted to compute the fractal dimension curve; its inflexion points were extracted whose Y-coordinates served as the image descriptor. These descriptors were submitted to the 1R algorithm, which ranks the three main measures. The LADtree was used as classifier, and finally the mass classification was suggested. The following sections provide a more detailed description of each step.

A. Image Dataset

A dataset of 250 breast masses was collected from the DDSM database (Digital Database for Screening Mammography) [18], in which 99 images were benign and 151 were malignant masses. An example of each type of mass is presented in Fig. 12.

(a) (b)

Fig. 2. Example of (a) benign and (b) malign masses.

B. Mass detection and edge extraction

In order to access the regions of interest (ROI), we applied an improved version of the EM/MPM algorithm [19]. This algorithm segments images, using techniques that combine a Markov Random Field and a Gaussian Mixture Model to obtain texture-based segmentation.

A fixed number of regions for segmentation had to be defined. In this work, we used five regions, as described in [9] which allowed satisfactory masks to be extracted for the contour extraction step, as shown in Fig. 3 (a), (b) e (c).

(a) (b) (c) (d)

Fig. 3. Example of the mass detection process using an improved EM/MPM. (a) Original image, (b) image segmented in five regions, (c) mass detected, (d) edge detection and extraction.

Next, the contour/edge from each ROI was obtained using the Sobel high-pass filter. Sobel is a discrete differentiation operator which computes an approximation of the gradient of the image intensity function [20]. Its function is to allow abrupt changes of intensity to be perceived, which characterize the boundaries of objects and, therefore, the detection of their contours. An example of result of the Sobel application is presented in Fig. 3 (d).

C. The Bouligand-Minkowski method

To implement the Bouligand-Minkowski method we used the description of [13]. The concept of fractal dimensions using this method is based on the correlation between an object interface and the space that it occupies. This correlation is analyzed on the basis of area which the object occupies after successive dilations. The fractal dimension is computed using the ramp of the linear regression of the set of points from a graph log(r) x log$A$(r). (This procedure is also called Minkowski's Sausage.) The fractal dimension d_f is defined by

$$d_f \sim 2 - \frac{\log A(r)}{r} \quad (1)$$

where $A(r)$ is the total area represented by the sum of all points at distance r of the dilation area. We used the Euclidean distance transform to implement the exact dilation technique.

The exact dilations of this shape corresponds to the sequence of all successive dilations, without repetition, by treating circles with increasing radius as structuring elements. The distances defined by the incremental radius are called exact distances, and the order in which a specific exact distance occurs is henceforth called its respective distance index **k**. Thereby, for **k**=0, Δ=0; **k**=1, Δ=1; **k**=2, Δ= $\sqrt{2}$, and so on. The set of all the dilated shapes for each of the possible exact distances corresponds to the exact dilations of the original shape [13].

Bouligand-Minkowski's fractal dimension approach can be defined by Algorithm 1.

Algorithm 1: Bouligand-Minkowski method

Data: $P = \{(x, y) | f(x, y) \in Img\}$, where Img corresponds to the pixels of the contourn

Result: The dilation area A

1 Compute and sort the Eudclidean distances $E = \{0, 1, \sqrt{2}, \sqrt{3}, 2\sqrt{2}, ..., l\}$ such that $l \in D$, where $D = \{r | r = (i^2 + j^2)^{1/2}; i, j \in \mathbb{N}\}$;

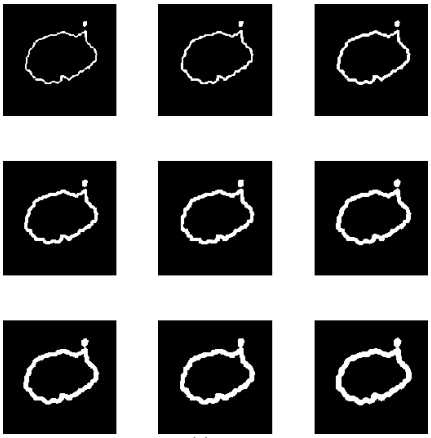
2 Compute $g_k(P)$ //the set of pixels at index distance k that belongs to E, so $g_0(P) = P$ and $g_k(P) = \{(x, y) | [(x - P_x)^2 + (y - P_y)^2]^{1/2} = E(k); x, y \in \mathbb{N}\}$;

3 Compute $Q(k)$ //the set of dilated pixels, defined by $Q(k) = \{(x, y) | g_k(P) - [g_k(P) \cap \bigcup_{i=0}^{k-1} g_i(P)]\}$;

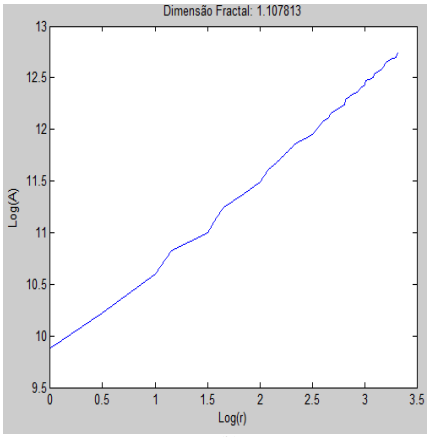
4 Compute $A(k) = \sum_{i=1}^{k} Q(i)$;

5 **return** A

Fig. 4 shows the stages to estimate fractal dimensions by means of Algorithm 1. Fig. 4(a) contains a representation of the distance transform, showing the dilations. Fig. 4(b) presents the bi-log curve log(r) x log$A(r)$, which was used to estimate the fractal dimension.



(a)



(b)

Fig. 4. (a) The first nine dilations using the Bouligand-Minkowski method, from top left to bottom right. (b) log(*r*) x log*A(r)* graph.

D. Multiscale analyses and feature vectors

Costa and Cesar Jr. [21] proposed a multiscale analysis of fractal dimensions which, unlike other methods of fractal estimation that generate a single number, reveals the variations of the fractal dimensions of an object based on the variations in the scale of the metric space occupied by it. It requires the log(*r*) x log*A'(r)* bi-log graph, where *A'* is the numerical derived point at (*r,A(r)*). Thus, in this work we use the Finite Differences method to compute *A'*.

Fig. 5 shows an example of the log(*r*) x log*A(r)* and log(*r*) x log*A'(r)* graphs.

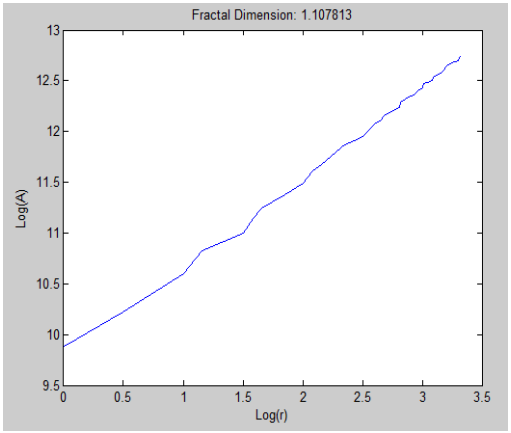
In this work, the Multiscale Fractal Dimension was calculated through the Differentiation by Finite Differences of the vectors of the areas of influence generated by the Bouligand-Minkowski method. How this new vector behaved along the changes in the radii of dilation in a 100*(r/max(r)) X log(*A'*) graph was then analyzed. From the multiscale signals, all twenty-two inflection points obtained with the radius of dilation previously specified were identified. Their coordinates on the Y-axis – log*A'(r)* – composed the feature vector.

E. Feature selection using 1R algorithm

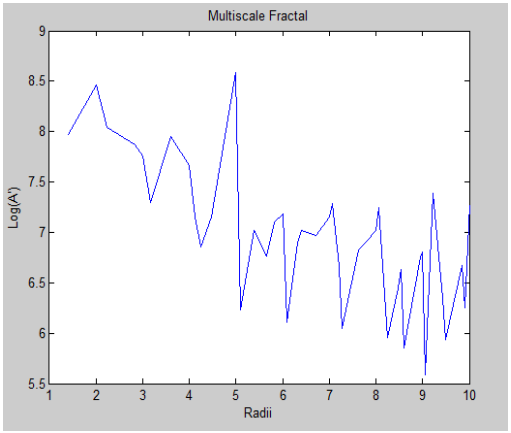
The feature vectors, which have 22 features, were submitted to an Attribute Worth Evaluator, based on the 1R algorithm [6].

OneR is a simple, but accurate, classification algorithm that generates a rule for each predictor in the data and then selects the rule with the smallest total error as its "one rule".

For our descriptor, 1R ranked all 22 attributes in order of relevance, highlighting the 11th, 15th and 17th points as the most relevant for classification. Thus, we considered the first three points to compose the feature vectors. A new descriptor was then formed, containing only these three measures as features.



(a)



(b)

Fig. 5. Multi fractal dimension method. (a) Fractal Dimension log(*r*) x log*A(r)* graph from the contour of the Fig. 3(d); (b) multiscale fractal dimension graph (100*(r/max(r)) X log(*A'*)).

F. Classification

For the classification task, we chose the LADtree classifier [7]. The LADTree is a classification technique that combines decision trees with the predictive accuracy of boosting into a set of interpretable classification rules, in this case using the LogitBoost strategy to directly induce the alternating decisions.

G. Performance evaluation of the proposed method

To validate the proposed method, we performed the experiments keeping well-known shape descriptors.

The first descriptor is composed of the 11 features extracted from the detected tumor: area, convex area, eccentricity, Euler number, extent, filled area, major axis length, minor axis length, orientation, perimeter and solidity.

The second well-known descriptor is called the Zernike moments. Its success is due to many desirable properties, such as rotation invariance, robustness to noise, expression efficiency, fast computation and multi-level representation for describing the shapes of patterns. We applied the Zernike moments and obtained vectors with 256 features.

To compare performance, we computed the accuracy, sensitivity and specificity measures. Accuracy denotes the percentage of predictions that are correct. Sensitivity is the measure of the ability of a prediction model to select instances of a certain class from a data set. The specificity corresponds to the true negative rate which is commonly used in two class problems. Accuracy, sensitivity and specificity are calculated, according to [22, 23], by using Equations 2, 3 and 4, respectively, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

An optimal prediction can achieve 100% sensitivity and 100% specificity. In our experiments, we considered values between 0 and 1.

We used the holdout approach, devoting 75% of the images of the dataset to training and the remainder to testing. We also used the k-fold cross-validation approach, in which the dataset is divided into k folds, a classifier is learned using k-1folds, and an error value is calculated by testing the classifier in the remaining fold. Finally, the estimation of the error is the average value of the errors committed in each fold. Thus, the error estimator depends on two factors: the training set and the partition into folds. In this work we chose k=10.

IV. Results

Tables I and II compare the accuracy, sensitivity and specificity of the proposed method and the other descriptors, all of which were submitted to the LADtree classifier, and takin the holdout and the 10-fold cross validation approaches, respectively.

TABLE I. Comparison of descriptors using the LADTree, Holdout Approach

Methods	Results		
	Precision	*Sensitivity*	*Specificity*
Multiscale Fractal	**0.8387**	**0.8148**	0.8571
Zernike	0.5806	0.4074	0.7142
Shape	0.7580	0.5555	**0.9142**

TABLE II. Comparison of descriptors using LADTree, 10-fold Cross-Validation Approach

Methods	Results		
	Precision	*Sensitivity*	*Specificity*
Multiscale Fractal	**0.7760**	**0.7272**	**0.8079**
Zernike	0.6560	0.5252	0.7417
Shape	0.7240	0.5454	**0.8079**

In Table I, we note that the proposed method (Multiscale Fractal) achieved the highest value of precision. Our method was 44% better than Zernike moments and 11% better than the Shape descriptor, taking the holdout approach.

Regarding the sensitivity and specificity values, all the three descriptors were more specific than sensitive. Our method presented the best sensitivity value, being 100% better than Zernike and 46% better than the Shape descriptor. The Shape descriptor achieved the highest value of specificity, being just 7% better than the proposed method. However, a better balance between sensitivity and specificity values was obtained by the Multiscale Fractal descriptor. That is to say, this method classified both malign and benign masses almost in the same proportion. However, the Shape descriptor presented wide difference between these values, which implies that most of the tumors were classified as benign. This diagnostic suggestion is dangerous when we consider that a specialist doctor might consequently not provide adequate treatment to malign cases classified as benign.

In Table II, with regard to the 10-fold cross validation approach, the Multiscale Fractal descriptor also provided the best values of precision and sensibility, being 18% more precise than Zernike moments and 7% more than the Shape descriptor. Although the precision values were not drastically greater, as in the previous approach, the difference between the sensibility values was still significant: between 33% and 38% compared to the other methods. The specificity values of our method and Shape descriptor were the same, but the Multiscale Fractal descriptor presented the best balance between sensibility and specificity, while Shape was 48% more specific than sensitive.

It is clear that the Multiscale Fractal descriptor achieved the highest values of accuracy, sensitivity and specificity when compared to Zernike and Shape. Furthermore, the proposed method shows a proper balance between sensitivity and specificity, unlike the other two descriptors.

V. Conclusion

In this work we presented a new method to characterize masses from mammograms, in which a multiscale fractal analysis was used to extract features, combining the 1R algorithm for feature selection and the LADtree classifier.

The low number of features deals with the *curse of high dimensionality* challenge. As image descriptor, it has only three features, while Zernike has 256 and Shape has 11 attributes.

Higher values of accuracy, specificity and sensitivity than either Zernike moments or values of Shape, and the balance between sensitivity and specificity values, show that our method reduces the *semantic gap*.

The results presented in this work indicate that our approach is well-suited for the task of classifying masses.

REFERENCES

[1] INCA, "Estimate 2016: cancer incidence in Brazil," Instituto Nacional de Câncer José Alencar Gomes da Silva, Rio de Janeiro2015.

[2] H. Li, X. Meng, T. Wang, Y. Tang, and Y. Yin, "Breast masses in mammography classification with local contour features," *BioMedical Engineering OnLine,* vol. 16, pp. 1-12, 2017.

[3] M. J. Silverstein, A. Recht, M. D. Lagios, I. J. Bleiweiss, P. W. Blumencranz, T. Gizienski, S. E. Harms, J. Harness, R. J. Jackman, and V. S. Klimberg, "Image-detected breast cancer: state-of-the-art diagnosis and treatment," *Journal of the American College of Surgeons,* vol. 209, pp. 504-520, 2009.

[4] J. Y. Choi, D. H. Kim, K. N. Plataniotis, and Y. M. Ro, "Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography," *Expert Systems with Applications,* vol. 46, pp. 106-121, 2016.

[5] R. M. Rangayyan and T. M. Nguyen, "Fractal analysis of contours of breast masses in mammograms," *J Digit Imaging,* vol. 20, 2007.

[6] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine learning,* vol. 11, pp. 63-90, 1993.

[7] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, and M. Hall, "Multiclass alternating decision trees," in *European Conference on Machine Learning*, 2002, pp. 161-172.

[8] S. Sharma and P. Khanna, "Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM," *Journal of Digital Imaging,* vol. 28, pp. 77-90, 2015.

[9] C. Y. V. Watanabe, M. X. Ribeiro, C. Traina, and A. J. M. Traina, "SACMiner: A New Classification Method Based on Statistical Association Rules to Mine Medical Images," in *Enterprise Information Systems: 12th International Conference, ICEIS 2010, Funchal-Madeira, Portugal, June 8-12, 2010, Revised Selected Papers*, J. Filipe and J. Cordeiro, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 249-263.

[10] S. P. Singh and S. Urooj, "An Improved CAD System for Breast Cancer Diagnosis Based on Generalized Pseudo-Zernike Moment and Ada-DEWNN Classifier," *Journal of Medical Systems,* vol. 40, pp. 1-13, 2016.

[11] S. Venkatalakshmi and J. Janet, "Classification of Mammogram Abnormalities Using Pseudo Zernike Moments and SVM," *International Journal of Image, Graphics and Signal Processing,* vol. 9, p. 30, 2017.

[12] M. A. Nogueira, P. H. Abreu, P. Martins, P. Machado, H. Duarte, and J. Santos, "Image descriptors in radiology images: a systematic review," *Artificial Intelligence Review,* vol. 47, pp. 531-559, 2017.

[13] O. M. Bruno, R. de Oliveira Plotze, M. Falvo, and M. de Castro, "Fractal dimension applied to plant identification," *Information Sciences,* vol. 178, pp. 2722-2733, 2008.

[14] M. Abdel-Nasser, J. Melendez, A. Moreno, O. A. Omer, and D. Puig, "Breast tumor classification in ultrasound images using texture analysis and super-resolution methods," *Engineering Applications of Artificial Intelligence,* vol. 59, pp. 84-92, 2017.

[15] R. Dobrescu, L. Ichim, and D. Crisan, "Diagnosis of breast cancer from mammograms by using fractal measures," *International Journal of Medical Imaging,* vol. 1, pp. 32-38, 2013.

[16] S. M. A. Beheshti, H. AhmadiNoubari, E. Fatemizadeh, and M. Khalili, "An Efficient Fractal Method for Detection and Diagnosis of Breast Masses in Mammograms," *Journal of Digital Imaging,* vol. 27, pp. 661-669, 2014.

[17] K. Falconer, *Fractal geometry: mathematical foundations and applications*: John Wiley & Sons, 2004.

[18] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th International Workshop on Digital Mammography*, 2000, pp. 212-218.

[19] A. G. Balan, A. J. Traina, C. Traina, and P. M. Azevedo-Marques, "Fractal analysis of image textures for indexing and retrieval by content," in *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*, 2005, pp. 581-586.

[20] U. G. Nevagi, A. Shahapurkar, and S. Nargundkar, "Edge detection techniques: a survey," *International Journal of Innovative Research and Development,* vol. 5, pp. 274-281, 2016.

[21] L. d. F. Costa and R. M. Cesar Jr., *Shape Analysis and Classification: Theory and Practice*, 2nd ed.: CRC Press, 2009.

[22] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: simple definition and calculation of accuracy, sensitivity and specificity," *Emergency,* vol. 3, pp. 48-49, 2015.

[23] S. Aruna, S. Rajagopalan, and L. Nandakishore, "Knowledge based analysis of various statistical tools in detecting breast cancer," *Computer Science & Information Technology,* vol. 2, pp. 37-45, 2011.