Received: 28 September 2019 | Revised: 25 February 2020 | Accepted: 10 March 2020

DOI: 10.1111/coin.12320

SPECIAL ISSUE ARTICLE

Computational Intelligence WILEY

Towards an approach using grammars for automatic classification of masses in mammograms

Ricardo Wandré Dias Pedro[1] | **Ariane Machado-Lima**[2] | **Fátima L. S. Nunes**[1,2]

[1]Electrical Engineering, Polytechnic School, University of São Paulo, São Paulo, Brazil

[2]Information Systems, School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, Brazil

Correspondence
Ricardo Wandré Dias Pedro, Electrical Engineering, Polytechnic School, University of São Paulo, São Paulo, Brazil.
Email: rwandre@usp.br

Abstract

Approximately 15% of all cancer deaths among women worldwide is due to breast cancer. Mammography is one of the most useful methods for the early detection of this disease. Over the last decade, several papers were published reporting the usage of different computer-aided diagnosis systems using pattern recognition techniques as a second opinion to obtain a more accurate diagnosis. However, the theory of formal languages has not been explored in this field. In this context, the main contribution of this study is to present the usage of a new syntactic approach that is able to classify breast masses found in mammograms as benign or malignant. The experimental tests were performed using a dataset that contains 111 images from different sources. The grammar-based classifiers achieved accuracy values ranging from 89% to 100% depending on the features and the model employed. Furthermore, to achieve a feature dimension reduction, a feature selection technique based on the Gini importance of each feature was employed. Additionally, we compared the obtained results with the grammar-based classifiers to the more traditional classifiers used in this research area, such as artificial neural networks, support vector machines, k-nearest neighbors, and random forest. The best result achieved by the grammar-based classifiers was approx-

imately 10% higher, in terms of accuracy, than the best results produced by the traditional classifiers, showing the strength of this grammatical approach.

KEYWORDS

breast cancer diagnosis, grammars, machine learning, mammography, mass classification, syntactic analysis

1 | INTRODUCTION

As reported by the World Health Organization (WHO)*, breast cancer is the most frequent cancer impacting women around the world and the cause of the greatest number of cancer-related deaths. It is estimated that this disease impacts more than 2 million women every year and, in 2018 alone, more than 627 000 women died from this type of cancer. Additionally, according to WHO, breast cancer rates are increasing globally in nearly every region. In consonance with WHO, Surveillance, Epidemiology, and Results Program (SEER)† estimated 268 600 new cases of this disease, which represents 15.2% of all new cancer cases, only in the United States for the year of 2019. Furthermore, as stated by SEER, the number of estimated deaths in 2019 is more than 41 000 (6.9% of all cancer deaths) in the North American society.

Mammography is one of the most reliable and effective methods for early-stage breast cancer detection.[1] Despite the fact that some rules can be applied to discern benign and malignant cases, only around 15% to 30% of the surgical biopsies are malignant.[2] An unnecessary biopsy can lead to problems, for instance, the physical pain the women are submitted to, the stress and anxiety until the diagnosis is confirmed and the financial cost of the procedure.[3,4]

Over the last decades, various computer-aided detection and computer-aided diagnosis systems have been proposed to detect and classify findings in mammograms, respectively. Several approaches of pattern recognition techniques have been developed and employed by the scientific community and industrial companies, as for instance, in the studies 5-8. The three most common machine learning techniques used to discriminate breast masses are artificial neural networks (ANN), support vector machines (SVM), and k-nearest neighborhood (KNN).[9] Albeit the theory of formal languages are being used to understand the content of images, including medical images,[10] the syntactic approaches are barely used to deal with mass classification.[9] Syntactic approaches have advantages such as dealing with hierarchical structures and their relationships found in images and the possibility to choose different type of grammars with different representation power at a cost of increasing model's complexity. These approaches are particularly attractive for problems where the data samples are not numerous as required by more recent techniques such as deep learning.[11]

To the best of our knowledge, besides our research group, only Tahmasbi et al[12] published a study in which grammars were employed in the process of mass discretization. Moreover, we could not find any paper describing the use of syntactic approaches to classify masses without combining grammars with other machine learning techniques.

*World Health Organization: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

†SEER Program: https://seer.cancer.gov/statfacts/html/breast.html

This paper expands a previous study showing preliminary results of a new syntactic approach for mass classification.[13] More precisely, the main contributions of the previous paper were the use of stochastic context-free grammars created based on the structures of AND-OR graphs to perform the discrimination of masses as benign or malignant.

The main contributions added in the present paper are:

1 using grammars to classify masses as benign or malignant, including: (i) parameter calibration process for feature discretization using the Omega algorithm;[14] (ii) testing the KBinsDiscretization algorithm for feature discretization;[15] (iii) using the concept of Gini importance for feature dimension and noise reduction; (iv) using a maximum posteriori algorithm to convert the deterministic grammars into stochastic grammars; (v) evaluation of the effect of dataset combination on the classifier learning process.
2 a fair comparison among classifiers created using grammars and classifiers created using the most common machine learning techniques (ANNs, SVMs, KNN, and random forest, RF) in the context of mass classification. For each traditional classifier, all possible feature combinations were tested in order to achieve the highest possible accuracy for each classifier.

2 | BACKGROUND

2.1 | Mammography

Mammography is an exam that uses X-rays to project the three-dimensional structures of a female breast in a two-dimensional (2D) image.[5] The exam was introduced in 1963 and brought an important revolution in breast cancer detection and treatment.[16] Mammography is an effective aid for early detection of breast cancer, since it is able to detect abnormal growth of tissues in the breast before they become palpable.[5] During a mammographic exam, normally, two projections of the breast are recorded[17]—the craniocaudal (CC) and the mediolateral oblique (MLO) views. Studies show that using the two views improves performing the detection and diagnosis of breast cancer when compared to using a single view.[17]

In general, women over the age of 45 are encouraged to undergo mammography screening at least once every two years, since the incidence of this disease tends to increase with age. Nevertheless, it is also common to find cases where young women were diagnosed with breast cancer.[16]

2.2 | Grammars

A grammar is a quadruple $G = (V_N, V_T, R, S)$, where V_N is a set of nonterminal symbols, V_T is a set of terminal symbols, R is a set of production rules and $S \in V_N$ is the initial symbol of the grammar. Considering the Chomsky hierarchy,[18] a context-free grammar is a type of grammar where the production rules are as follows:

- $A \rightarrow \beta, A \in V_N, \beta \in (V_T \cup V_N)^*$

A stochastic grammar is a quintuple $G_s = (V_N, V_T, R, S, P)$, where:

- V_N, V_T, S, have the same meaning as described previously;

- $R \subset \{\alpha \rightarrow \beta, p\}$, where $\alpha \rightarrow \beta$ is a production and $p, 0 \leq p \leq 1$, is a probability associated to the production;
- For each α that is the left side of a production rule, considering all productions $\{\alpha \rightarrow \beta_i, p_i\} \in R, \beta_i \in (V_N \cup V_T)^*$, we have $\sum_{i=1} p_i = 1$;
- P is the set of probabilities of the production rules.

A parser for a given grammar G is an algorithm that, given a sequence x, is able to provide one or all the syntactic trees for that sequence if it belongs to the language generated by G, otherwise, the parser generates an error. For a stochastic grammar G_s, the parser not only generates the syntactic trees, but also provides the probability of x according to the grammar G_s.

Stochastic grammars are useful to classify a new object, since it is just necessary to compute the probability of that object belonging to a class by computing a parse tree, and then choosing the class with the highest probability.

3 | RELATED WORKS

Numerous works have dealt with the problem of mass classification in digital mammograms.[9] Among these works, the studies[19-23] were conduced using the same set of images and features used in the present work.

The studies[19,20] presented the use of concavity and convexity fractions combined with compactness and spiculation index as features to serve as input to a discriminant analysis program for the classification of masses. The authors achieved an accuracy of 81% in Reference 19 for the classification of masses as benign or malignant and an accuracy of 91% in Reference 20 when classifying masses as circumscribed or spiculated.

The study[21] showed the use of genetic programming and feature selection in the mass classification processes. The authors used shape and texture features extracted from 57 images from the Alberta Program for the Early Detection of Breast Cancer database (Screen Test),[24] obtaining a specificity of 95% and a sensitivity of 97.3%.

The features compactness, spiculation index, fractal dimension, and fractional concavity were employed in Reference 22. In the classification process, the higher value of the area under the Receiver Operating Characteristic (ROC) curve was 0.92 with images combined from Mammographic Image Analysis Society database (MIAS)[25] and from Screen Test database.

In Reference 23, gradient and texture features were combined. To feed the classifiers based on posterior probabilities computed from Mahalanobis distance, the authors provided features based on gray level co-occurrence matrices, acutance, and coefficient of gradient strength variation. The higher value of the area under the ROC curve was 0.76 with combined images from MIAS and Screen Test databases.

The work 12 demonstrated the use of syntactic approach as part of the process of mass classification. In the work, the authors used the output of a syntactic analysis as input to an ANN classifier. Features based on texture and Zernike moments were extracted from images coming from MIAS database. As a result, the average area achieved under the ROC curve was 0.861.

More recently, other approaches were employed in the process of mass classification. For instance, in study 26 the use of an ANN considering the four BI-RADS categories was reported. The authors used texture and shape features together with patient age as input to the classifier. The experiments were performed using 480 mass images selected from DDSM database[27] and the highest accuracy rate achieved was 88.02% considering benign and malignant classes.

A convolutional neural network (CNN) was employed in Reference 28. The authors used 600 images selected from DDSM database achieving an accuracy rate of 97.4%.

RF and CNN were used in the study 8 to solve the problem of mass classification. The best results were accuracy of 0.95 with the RF on the CNN with pretraining, accuracy of 0.91 with the CNN with pretraining and accuracy of 0.90 with the RF on hand-crafted features.

In the work 29 a novel representation of shape masses using a sparse region of interest was presented. Several features were extracted from this new representation using gray level co-occurrence matrix (GLCM) and gray level aura matrix to serve as input to a multi-SVM. The classification accuracy was of 97.2% using 322 images from MIAS database and considering the following classes: calcifications, circumscribed, spiculated, ill-defined, architectural distortions, asymmetry(s), and normal.

Muramatsu et al[30] proposed the use of radial local ternary patterns as features to discriminate masses as benign or malignant. They tested their approach using ANN, SVM, and RF. Moreover, they compared the classification ability of the proposed feature against regular local ternary pattern, rotation invariant uniform local ternary pattern, texture features based on the GLCM and wavelets features. The highest area under the ROC curve was of 0.90 achieved with the proposed features and the ANN classifier.

4 | MATERIALS AND METHODS

4.1 | Database

The first dataset (dataset 1) contains 54 images from MIAS database and from the teaching library of the Foothills Hospital in Calgary.[20] Images from MIAS have spatial resolution of 50μm, while images from Foothills Hospital have spatial resolution of 62μm. The other dataset (dataset 2) has 57 images from Screen Test database.[24] The images have a spatial resolution of 50μm, but to extract gradient/texture features the images were resampling to 200 μm.[21]

The boundaries of all masses were manually drawn by specialists[22] and the 111 images were labeled with the following labels: (i) circumscribed benign (CB); (ii) spiculated benign (SB); (iii) circumscribed malignant (CM); and (iv) spiculated malignant (SM). The benign and malignant labels were used as class labels in the classification process, while the circumscribed and spiculated labels were used as shape labels. Considering the two datasets combined, there are 53 CB masses, 10 CM masses, 12 SB masses, and 36 SM masses. Figure 1 shows the distribution of the masses considering their labels and the dataset.

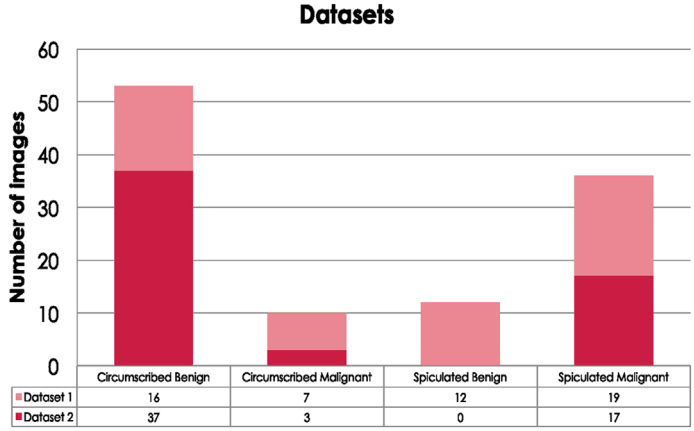


FIGURE 1 Distribution of masses according to their labels. From 65 benign masses, 81.5% are circumscribed and 18.5% are spiculated. From 46 malignant masses, 21.7% are circumscribed and 78.3% are spiculated [Color figure can be viewed at wileyonlinelibrary.com]

4.2 | Features

To deal with the mass classification problem, a total of 12 features were extracted from gray-level images belonging to the datasets. Eight features represent the shape characteristics of the masses, three features are related to the image gradient and one feature represents the mass texture. These same features were previously extracted and used in the studies.[19-23]

4.2.1 | Shape features

Compactness:
The feature compactness (CC) is useful to verify the efficiency of a contour to cover a given area.[22] This measure was obtained considering Equation (1),[20] where P is the perimeter of the mass and A is its area.

$$\mathrm{CC} = 1 - \frac{4\pi A}{P^2}. \tag{1}$$

Spiculation index:
The spiculation index (SI) is a measure used to verify how spiculated a nodule is. It can be obtained by Equation (2),[20] where S_i and θ_i for $i = 1, 2, 3, \ldots, N$ represent the length and the angle of N sets of segments corresponding to N mass spicules, respectively.

$$\mathrm{SI} = \frac{\sum_{i=1}^{N}(1 + \cos\theta_i)S_i}{\sum_{i=1}^{N} S_i}. \tag{2}$$

Fractional concavity:
Fractional concavity (FC) is a measure based on the number of concavities in a mass boundary. This measure is useful, since benign masses tend to have more convex parts and malignant masses tend to have both concave and convex parts. Let $S_i, i = 1, 2, 3, \ldots, M$, be the size of M segments from a mass boundary. The total length of the boundary T_l is computed using Equation (3). Let $\mathrm{CC}_i, i = 1, 2, 3, \ldots P$, be the length of P concave segments, the length of all concave segments CC_l is given by Equation (4). The FC is computed by Equation (5).[19,20]

$$T_l = \sum_{i=1}^{M} S_i. \tag{3}$$

$$\mathrm{CC}_l = \sum_{i=1}^{P} \mathrm{CC}_i. \tag{4}$$

$$\mathrm{FC} = \frac{\mathrm{CC}_l}{T_l}. \tag{5}$$

Fractal dimension:
Fractal dimension (FD) is a measure used to verify self-similarity, nested complexity or the capacity of filling spaces of a shape. In general, this measure can be used to explain the complexity of

a pattern.[22] A self-similarity D is defined by Equation (6), where a is the number of self-similar pieces considering a reduction factor of $1/s$, which can be obtained by Equation 7. Thus, the slope of a straight-line approximation of a plot $\log(a)$ vs $\log(1/s)$ can be considered as an estimation of D.[22]

Four different measures of FD were used in this study. The first two were fractal dimensions computed using the 2D contour of the masses applying the methods *ruler* and *box counting*.[22] The other two fractal dimension measures were calculated considering the one-dimensional (1D) signature of the contour of the masses also using the methods *ruler* and *box counting*.

$$D = \frac{log(a)}{log(1/s)}. \tag{6}$$

$$a = \frac{1}{s^D}. \tag{7}$$

Fourier Factor:
The Fourier Factor (FF) is a measure related to the roughness or the presence of high-frequency components in a contour.[22] This measure can be obtained by Equation 8, where $Z_0(k)$ are the Fourier descriptors normalized (Equation (9)), $Z(k)$ are the Fourier descriptors (Equation (10)), for $k = -N/2, \dots, -1, 0, 1, 2, \dots N/2 - 1$, and $z(n) = x(n) + jy(n), n = 0, 1, \dots N - 1$ is the sequence of contour pixels.[22]

$$\text{FF} = 1 - \frac{\sum_{K=-N/2+1}^{N/2} |Z_0(k)|/|k|}{\sum_{K=-N/2+1}^{N/2} |Z_0(k)|}. \tag{8}$$

$$Z_0(k) = \begin{cases} 0, & k = 0; \\ \frac{Z(k)}{|Z(1)|}, & \text{otherwise.} \end{cases} \tag{9}$$

$$Z(k) = \frac{1}{N} \sum_{n=0}^{N-1} z(n) exp\left[-j\frac{2\pi}{N}nk\right]. \tag{10}$$

4.2.2 | Gradient/texture features

Acutance:
Acutance (A) is a measure of the change in density across a mass boundary.[23] It is computed using directional derivatives along a line of pixels in the normal direction in each point of the mass boundary. This feature was calculated considering Equation 11, where $f_{\max}$ and $f_{\min}$ are the local maximum and minimum pixel values along the line of pixels in the normal direction in each point of the mass boundary, respectively, and d_i is the root-mean-squared gradient at the ith boundary point.

$$A = \frac{1}{f_{\max} - f_{\min}} \frac{\sum_{i=1}^{N} d_i}{N}. \tag{11}$$

Two versions of acutance were used in this study. The first one (traditional acutance [TA]), was computed considering the difference between pairs of pixel values along the normal direction of the mass boundary. The second one (acutance), was calculated in a similar way, but considering the differences between pairs of adjacent pixels along the normal direction of the mass boundary.[23]

Coefficient of variation:
The aim of coefficient of variation (CV) is to investigate how the sharpness of a mass varies around its boundary, besides evaluating its average sharpness with the measure of acutance.[23] The value of the CV can be obtained using Equation (12), where $M = 5, f_i(n), n = 0, 1, 2, \ldots, n_i$ are the pixels considered in the ith boundary point considering the perpendicular direction, and μ_ω is defined according to Equation (13).[23]

$$\sigma_\omega^2 = \frac{1}{M} \sum_{n=[-M/2]}^{[M/2]} [f_i(n) - \mu_\omega]^2. \tag{12}$$

$$\mu_\omega = \frac{1}{M} \sum_{n=[-M/2]}^{[M/2]} f_i(n). \tag{13}$$

Contrast:
A GLCM that considers the pixels of the mass boundary and their surroundings was used to extract the feature contrast (CO). A GLCM $P_d(i, j, \theta, d)$ reflects the probability distribution of the transition of gray-level i to a gray-level j, considering the direction θ and a distance d (in this case, $d = 1$). The contrast measure used in this study is given by Equation (14), where N is the number of gray-levels (256 in this study) and R is the total pairs of pixels in the region used in the specified angular direction.[23]

$$\text{Contrast} = \sum_{n=0}^{N-1} n^2 \sum_{i-j=n} \left(\frac{P(i,j)}{R} \right)^2. \tag{14}$$

4.3 | Mass classification approach

The approach used in this study consists of the feature extraction, the grammar-based classifiers development (Sections 4.3.1 to 4.3.5) and a comparison among grammar-based classifiers and other classifiers, such as, ANN, SVM, KNN, and RF (Sections 6 and 7). The features used in the present work have been extracted by the researchers of studies 19-23. Figure 2 shows the pipeline of tasks that were executed in order to create and validate the proposed grammar-based classifier. In the next sections, each task is explained in detail.

4.3.1 | Feature selection

The RF classifier implicitly performs feature selection for the classification task. The output of the RF learning algorithm includes the "Gini importance" of each feature. The Gini importance of a

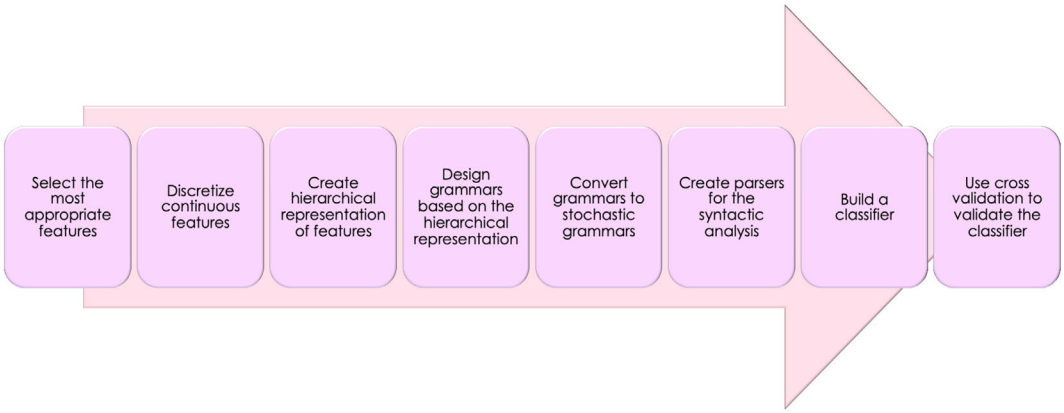


FIGURE 2 Pipeline of tasks executed to create and validate the proposed approach [Color figure can be viewed at wileyonlinelibrary.com]

feature measures how often this feature was selected for a split, and how important its discriminative value was for the classification process.[31] Selection of the most discriminative features is important not only to reduce dimensionality saving resources such as memory, CPU, and time of processing, but also to reduce noise in the data, since not all features necessarily contribute to increase the accuracy of the model. In fact, feature selection techniques have been applied to several pattern recognition fields, such as fingerprint and face recognition.[32]

This study used the Gini importance as a feature selection criterion to choose the most adequate features to create the proposed grammar models (first step in Figure 2).

4.3.2 | Feature discretization process

Omega algorithm[14] and KBinsDiscretizer algorithm[15] were employed in order to discretize the continuous features (second step in Figure 2). This task is necessary since each value of a feature has to be represented by a symbol or token in formal languages. Thus, these symbols can be used to create a sequence of symbols that will represent a benign or malignant mass. For instance, consider that the values {0.1, 0.2, 0.3, 0.4, 0.5} are the only possible values of compactness feature, and the values {0.5, 0.8, 1.0, 1.1} are the only possible values of spiculation index feature. A discretization process could label the compactness values 0.1, 0.2 as "*cc1*," and the compactness values 0.3, 0.4, and 0.5 as "*cc2*" ("*cc*" to specify the compactness feature and "*1*" and "*2*" to represent the two intervals/bins). Similarly, the process could label the instances 0.5 and 0.8 as "*si1*" and 1.0 and 1.1 as "*si2*" for the spiculation index.

Omega algorithm has two input parameters, $H_{\min}$ and $\zeta_{\max}$. The parameter $H_{\min}$ specifies the minimum number of elements for each bin, that is, the minimum number of elements that each symbol/token should represent. The parameter $\zeta_{\max}$ is the maximum inconsistency level, determining that two consecutive bins can be merged only if their elements have the same majority class and if the inconsistency level is below $\zeta_{\max}$.[14] In the present study we performed a calibration process in order to choose the best input parameters to handle the classification process.

The implementation of KBinsDiscretizer used in this work has three input parameters: *n_bins* is the number of bins to be generated, *encode* is the method used to encode the result, that is, how the bin identifier is generated, and *strategy* is the strategy used to define the widths of each bin.

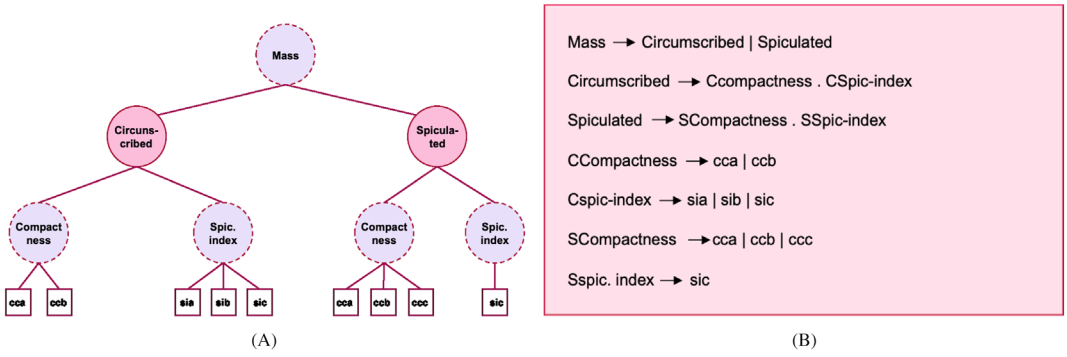
4.3.3 | Hierarchical models and grammars

AND-OR graphs are hierarchical models able to represent context-free languages and also some context dependencies.[33] Here, the AND-OR graphs were used to supply a visual representation of the features extracted from masses in a hierarchical form (third step in Figure 2). Since we have not included context dependencies in these representations, each AND/OR graph was converted into an equivalent context-free grammar. The AND/OR (internal) nodes can be seen as the non-terminal symbols of the grammar, and the leaf nodes can represent the terminal symbols of the grammar. The AND nodes are used to decompose entities into their parts (representing the right side of a production, composed of a concatenation of symbols), while the OR nodes are used to produce alternative substructures (ie, alternative productions for the same left side symbol).[33]

Three core hierarchical models were created to represent the masses considering several different combinations of shape and gradient/texture features. Each core model derives two models, one for benign and other for malignant masses, where the leaf nodes are the discretized values of the features from benign/malignant labeled images, respectively. The difference between a benign and malignant representation, considering these models, are due to the possible values of the features in the leaf nodes.

The first core model, named **shape-only model** (Figure 3A), considers that a mass can be circumscribed or spiculated and uses only shape features to discriminate masses as benign or malignant. In Figure 3A, two shape features are represented (compactness and spiculation index) together with the Circumscribed/Spiculated labels. The leaf nodes represent the values of these features and they were discretized using the labels CB, CM, SB, and SM. In addition, Figure 3B shows the grammar designed from this AND-OR graph (fourth step in Figure 2).

The second core model, named **shape-texture model**, is shown in Figure 4A and also considers that a mass can be circumscribed or spiculated. However, it uses not only shape, but also gradient/texture features in the process of masses discrimination as benign or malignant. In this model, the shape features (compactness and spiculation index) were discretized using the labels CB, CM, SP, and SM, while the gradient/texture features (contrast and acutance) were discretized using the labels Benign (B) and Malignant (M). Figure 4B shows the grammar designed from this AND-OR graph (fourth task in Figure 2).



(A) (B)

FIGURE 3 (A) Core hierarchical model (**shape-only model**) for representation of masses using AND-OR graphs. (B) The designed grammar. The solid circles are the AND nodes, the dashed circles are the OR nodes and the squares are the leaf nodes. The symbol "|" represents the logic condition OR and the symbol "." represents the logic condition AND (concatenation operation in formal languages) [Color figure can be viewed at wileyonlinelibrary.com]

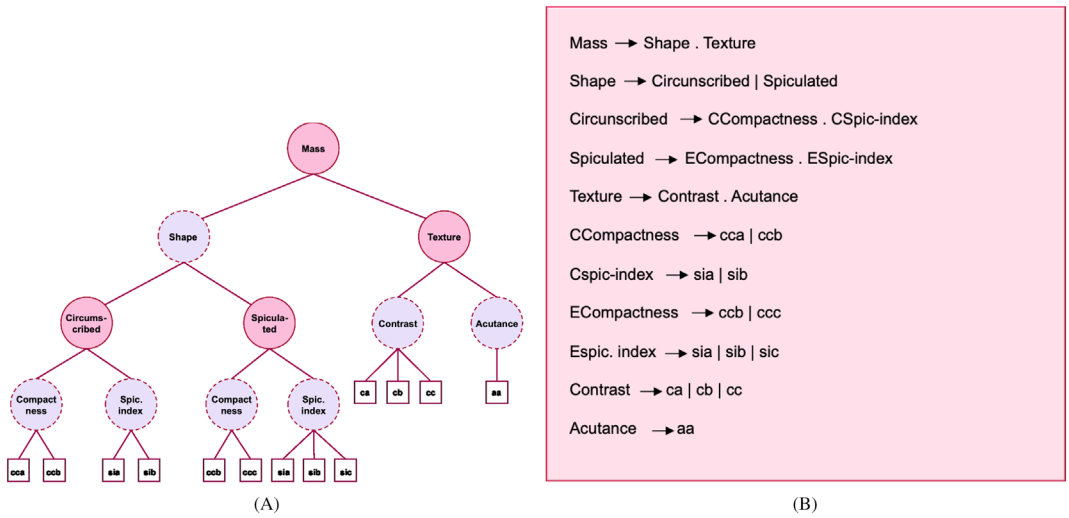


FIGURE 4 (A) Core hierarchical model (**shape-texture model**) for representation of masses using AND-OR graphs. (B) The designed grammar. The solid circles are the AND nodes, the dashed circles are the OR nodes and the squares are the leaf nodes. The symbol "|" represents the logic condition OR and the symbol "." represents the logic condition AND (concatenation operation in formal languages) [Color figure can be viewed at wileyonlinelibrary.com]

The last core model, named **no-shape-label model**, and its grammar are displayed in Figure 5A,B (fourth step in Figure 2), respectively. This model does not take into consideration whether a mass is circumscribed or spiculated, but it uses the shape and the gradient/texture features to classify masses as benign or malignant. In addition, the discretization process considered only the labels (B) and (M) for shape and gradient/texture features.

To convert the context-free grammars generated in the previous step in stochastic context-free grammars, we used a maximum posteriori algorithm proposed in Reference 34 (fifth step in

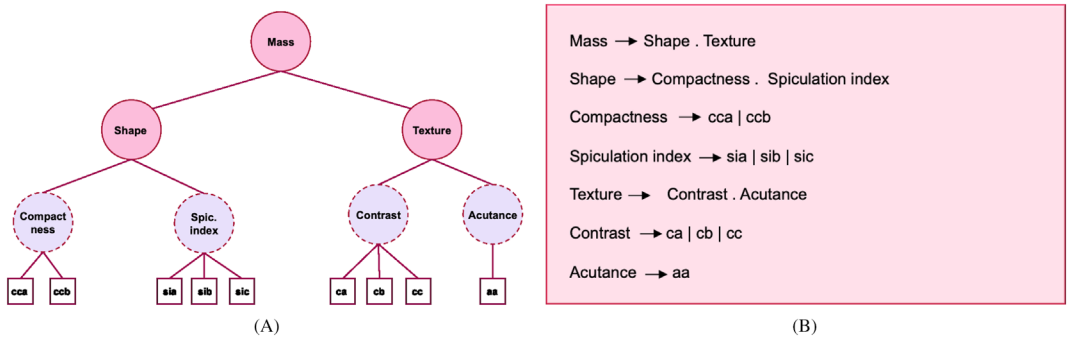


FIGURE 5 (A) Core hierarchical model (**no-shape-label model**) for representation of masses using AND-OR graphs. (B) The designed grammar. The solid circles are the AND nodes, the dashed circles are the OR nodes and the squares are the leaf nodes. The symbol "|" represents the logic condition OR and the symbol "." represents the logic condition AND (concatenation operation in formal languages) [Color figure can be viewed at wileyonlinelibrary.com]

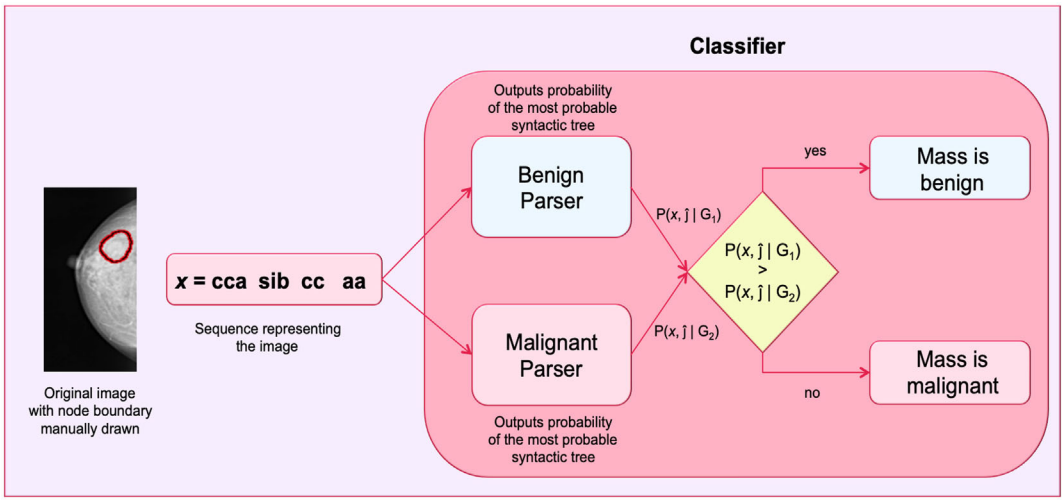


FIGURE 6 Process of a mass classification considering the benign and malignant classes [Color figure can be viewed at wileyonlinelibrary.com]

Figure 2). In this algorithm each production rule of the grammar has a counter, initialized[‡] with value 0.1. The next step is to use this initial grammar to parse each sequence in the training set and, every time a production rule is used during the parsing, its counter is incremented by 1. After analyzing all sequences in the training set, the counter values are normalized into a probability value so that the sum of probabilities of production rules with the same left side is equal to 1.

4.3.4 | The classifier

A version of the Earley algorithm[35] for stochastic grammars was used to parse the sequences representing the masses, considering the stochastic context-free grammars created (sixth step in Figure 2). Each built classifier (seventh step in Figure 2) contains a parser generated for the benign grammar (G_1) and another parser generated for the malignant grammar (G_2). For a given image x represented by its sequence, each parser i is used to provide $P(x, \hat{\eta}|G_i)$ where $\hat{\eta} = \text{argmax}P(x, j|G_i)$, that is, the probability value of the most probable syntactic tree of x given by grammar G_i, i=1,2. The mass image is classified as benign if $P(x, \hat{\eta}|G_1) > P(x, \hat{\eta}|G_2)$, otherwise it is classified as malignant. Figure 6 shows the process of a mass classification. In this example, the mass is represented by the sequence *cca sib cc aa*, which consists of the discretized values of the features compactness, spiculation index, contrast, and acutance. The parsers created that consider the grammars used to represent the benign and malignant masses, analyze the sequence and cast the most probable syntactic trees and their probabilities. The classifier then labels the mass as belonging to the class whose parser provided the highest probability.

[‡]Initialize the counters with 0.1 are important to guarantee that, at the end of the process, no production rule will have probability equals to zero.

4.3.5 | Validation

A k-fold cross validation technique ($k = 23$) was used to test the three core hierarchical models and their variations (different number of features) with the two datasets combined (eighth step in Figure 2). This value of k was chosen due to the small number of available images in the database, calculated in order to allow the classifier learning from 106 images and testing with five images, approximately, in each iteration (fold).

Moreover, to verify the performance of the classifiers, the metrics sensitivity, specificity, accuracy and the estimated area under the ROC curve (AUC) as proposed in Reference 36 were calculated.

5 | EXPERIMENTS AND RESULTS

This section presents the experiments and the results obtained by the classifiers based on grammars. During the feature discretization, when Omega algorithm was employed, we performed a calibration procedure varying the values of $H_{\min}$ (using 2, 3, 4, and 5) that specify the minimum number of elements that each symbol can represent and the values of $\zeta_{\max}$ (0.35, 0.40, and 0.45) that specify the maximum inconsistence level. For the KBinsDiscretizer algorithm, we varied the values of *n_bins* (using 20, 30, and 40) and keep the value of *encode* as "ordinal" and the value of *strategy* as "uniform" (then, the bins have all the same size). Thus, the impact of the discretization process on the mass classification can be analyzed. Moreover, the tests were performed combining the datasets described in Section 4.1.

In addition, the Gini importance of each feature was used as feature selection criterion during the creation of some models. To calculate the Gini importance a RF classifier was created using 70% of the images for training and the other 30% of the images for testing. The classifier (RandomForestClassifier) was created using the python sklearn.ensemble API[§]. Table 1 shows the Gini importance of each feature.

In the created models, except when all shape and texture features were used, we limited the number of each type of feature to 70 % of the total of that type of features. For instance, when the Gini importance was used to select the shape features, the highest number of features used was five of the eight possible features. For texture, when the Gini importance was considered, the highest number of features used was two of the four possible features.

5.1 | Shape-only model

The **shape-only model** is the core hierarchical model where only shape features are used. The model considers whether the masses are circumscribed or spiculated while the discretization process *when using Omega algorithm* considers the labels CB, SB, CM, and SM. It is important to mention that although these shape labels are used to create the model, they are not required during the classification task. The KBinsDiscretizer does not consider any label during the discretization process (Figure 3A in Section 4.3.3)

[§]**RandomForestClassifier:** https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

TABLE 1 Gini importance of each feature according to the random forest classifier. The features are ordered according to the Gini importance, from the highest to the lowest value

Feature	Type	Gini importance
Fractal dimension 1D Ruler (1R)	Shape	0.161724998
Spiculation index (SI)	Shape	0.135610343
Fractional concavity (FC)	Shape	0.129359555
Fourier factor (FF)	Shape	0.107758047
Fractal dimension 2D Ruler (2R)	Shape	0.094202521
Fractal dimension 1D Box counting (1B)	Shape	0.079181666
Compactness (CC)	Shape	0.066315130
Contrast (CO)	Texture	0.063269016
Fractal dimension 2D Box counting (2B)	Shape	0.056634071
Traditional acutance (TA)	Gradient	0,040501998
Acutance (AC)	Gradient	0.034178034
Coefficient of variation (CV)	Gradient	0.031264622

The results of the mass classification when Omega algorithm was employed can be seen in Tables A1, A2, A3, and A4. These tables show the accuracy when three, four, five, and all features were used to create the models (the features were selected according to the Gini importance). As can be noticed, with the increasing values of $H_{\min}$ the accuracies tend to decrease. The reason for this fact might be that the higher the $H_{\min}$, the classes belonging to the same bin in the discretization process can be more mixed. Moreover, it can be seen that the values of $\zeta_{\max}$ chosen to the discretization process had a minimum impact in the classification of masses, since the accuracies tend to be the same as the $\zeta_{\max}$ changes. Furthermore, the highest accuracies (0.97) were achieved when five and eight features were used and, the accuracies tend to decrease when the models used fewer features.

5.2 | Shape-texture model

The **shape-texture model** is one of the core hierarchical models where shape and texture features are used. The model considers whether the masses are circumscribed or spiculated and the discretization process when Omega algorithm is applied considers the labels CB, SB, CM, and SM for the shape features and the labels Benign (B) and Malignant (M) for the texture/gradient features. As previously (Section 5.1), the KBinsDiscretizer does not consider any label during the discretization process (Figure 4A in section 4.3.3).

Tables A1, A2, A3, and A4 show the results of the mass classification when five (three shape and two texture), six (four shape and two texture), seven (five shape and two texture), and when all eight shape and four texture/gradient features were used to create the model and Omega algorithm was used in the discretization process. Again, the features used were selected according to their importance given by the Gini coefficient. Analyzing these tables, it was observed that the accuracy values did not change too much when the number of features decreased. The highest accuracy obtained was one with the models with 12 features. Nevertheless, the values of $H_{\min}$

continue exerting a significant impact on the performance of the models, since when this parameter increases, the accuracies tend to decrease.

The results of the classification using KBinsDiscretizer as discretization algorithm are shown in Tables A5, A6, A7, and A8. This model achieved the maximum accuracy (one) when all the 12 features were used.

5.3 | No-shape-label model

The **no-shape-label model** is the core model where shape and texture features are used without considering the shape labels circumscribed and spiculated. The discretization process using Omega algorithm considered only the labels Benign (B) and Malignant (M) while the KBinsDiscretizer algorithm does not consider any labels (Figure 5A in Section 4.3.3).

The mass classification results considering the classes benign and malignant and Omega algorithm are shown in Tables A1, A2, A3, and A4. These tables display the results when five (three shape and two texture), six (four shape and two texture), seven (five shape and two texture), and all the eight shape and the four texture/gradient features were included to create the model. The best accuracy achieved was 0.93 considering four shape and two texture features. Yet, the $H_{\min}$ parameter had a major role in the classification process, since when this parameter increases, the accuracies tend to decrease.

Tables A5, A6, A7, and A8 show the results when the KBinsDiscretizer was used in the discretization process. The maximum accuracy (one) was obtained with eight shape and four texture features and the same pattern previously observed has repeated, that is, when the number of bins increases, it also increases the accuracy of the model.

6 | OTHER CLASSIFIERS

According to a systematic review previously conducted,[9] the most used pattern recognition techniques to handle mass classification are ANN, SVM, and KNN, in this order. In this study, these three techniques were implemented using the python sklearn library[15] in order to compare the accuracies achieved by the models based on grammar in contrast to the accuracies achieved by the most employed techniques. Moreover, the RF technique was also implemented using the sklearn library, as there are also numerous works that have used RF or decision trees to deal with the mass classification problem.[9]

For a fair comparison, all classifiers were evaluated using a paired cross-validation, where each fold contained exactly the same images for all the classifiers. Before executing the tests, all features were standardized and, consequently, their values ranged from 0 to 1. This step is important in order to avoid different influences of the feature scales.

Table 2 shows the hyperparameters used in the execution of each classifier. In addition, all possible combinations of the 12 features were tested, given a total of 4095 different feature subsets. This brute force approach was employed to guarantee that at the end of the tests, we would have the highest accuracy for each classifier considering the chosen hyperparameters.

The highest and lowest accuracies obtained by each classifier are shown in Tables A9 and A10. It is observed that the highest accuracies achieved by ANN, SVM, KNN, and RF were 0.92, 0.92, 0.90, and 0.89, respectively.

TABLE 2 Hyperparameters tested for each classifier. For artificial neural networks (ANN): α is the regularization term; *learning_rating* is the step size used to update the weights; *n_neurons* is the number of neurons in the hidden layer (only one hidden layer was used); and *f_activation* is the activation function. For support vector machines (SVM): *kernel* is the type of kernel used; C is the penalty parameter of the error term; γ is the kernel coefficient. For k-nearest neighborhood (KNN): k is the number of neighbors. For random forest (RF): *n_estimators* is the number of threes used; *max_features* is the number of features used when searching for the best split

Classifier	Hyperparameters
ANN	$\alpha = 0.0001$;
	learning_rating $= 0.001, 0.01, 0.1, 1$;
	n_neurons $= 2, 3$;
	f_activation = sigmoid, hyperbolic tangent, linear
SVM	*kernel* = linear, polynomial, radial basis function;
	$C = 0.01, 0.1, 1, 5, 10, 50, 100$;
	$\gamma = \frac{1}{n\_features}$
KNN	$k = 1, 3, 5, 7, 9$
RF	*n_estimators* $= 100$;
	max_features $= \sqrt{n\_features}$

7 | DISCUSSION

As presented in Section 5, various experiments were performed to verify if masses could be correctly classified as benign or malignant using grammars. More precisely, three core hierarchical models were created using several variations of features and two algorithms were used to discretize features to verify the robustness of the proposed approach.

The **shape-only model** used only shape features and the highest accuracies achieved when Omega algorithm was used were of 0.97, 0.97, 0.96, and 0.91 when eight (all), five, four, and three features were used and $H_{\min} = 2$, respectively. Its worst accuracies achieved considering the same number of features were 0.89, 0.91, 0.92, and 0.92 with $H_{\min} = 5$. When KBinsDiscretizer was used, the best accuracies were 0.97 (eight and five features), 0.95 (four features) and 0.94 (three features) using *n_bins* $= 40$, while the worst accuracies were 0.94 (eight features), 0.92 (five and four features), and 0.91 (three features) using *n_bins* $= 20$.

The **no-shape-label model** used shape and texture features and does not consider whether the masses are circumscribed or spiculated. The highest accuracies achieved when Omega algorithm was used were 0.93, 0.92, 0.93, and 0.92 with $H_{\min} = 2$ when 12 (all), seven, six, and five features were used, respectively. The worst accuracies obtained by this model were of 0.90, 0.88, 0.90, and 0.91 when the same number of features were used, but with $H_{\min} = 5$. Using KBinsDiscretizer as discretization method, the best accuracies were 1, 0.97, 0.95, and 0.92 when 12 (all), seven, six and five features were used, respectively. Yet, the worst results were accuracies of 0.92, 0.90, 0.89, and 089 considering the same number of features.

Overall, when Omega was used the highest accuracies were achieved by the **shape-texture model**. This is the model that uses shape and texture features and considers whether the masses are circumscribed or spiculated. The accuracies were of 1, 0.99, 0.99, and 0.98 when the $H_{\min} = 2$

and 12 (all), seven, six, and five features were used, respectively. The worst accuracies obtained by this model, considering the same number of features were, respectively, 0.94, 0.95, 0.95, and 0.93 with $H_{\min} = 5$ or $H_{\min} = 4$. When KBinsDiscretizer was employed, the best accuracies were 1, 0.99, 0.96, and 0.96 considering 12 (all), seven, six, and five features. The worst accuracies were of 0.97, 0.94, 0.92, and 0.93 for the same number of features.

Figure 7 exhibits the highest accuracies by each core hierarchical model when the number of features changed and Omega algorithm was used. Note that the **shape-only model** uses fewer features than the other models, since this model does not use the gradient/texture features. As can be seen, the **shape-only model** is the model that has a bigger impact when the number of features decreases (accuracy goes from 97% with eight shape features to 91% with three shape features). The other two core hierarchical models proved to be more stable when the number of features decreased.

Figure 8 shows the highest accuracies by each core hierarchical model when the number of features changed and KBinsDiscretizer algorithm was used. In general, it can be noticed that when all features are used, the accuracies are higher than the ones achieved when Omega algorithm was used. Nevertheless, as the number of features decreases, the accuracy tends to decrease more abruptly.

In general, one possible reason the **shape-only model** is less robust could be the absence of texture features. Usually, the circumscribed masses tend to be benign while the spiculated ones tend to be malignant. Consequently, the shape features are ideal to be used in the classification process. Nevertheless, some benign masses can be spiculated and some malignant masses can be circumscribed and, in this scenario, the absence of texture features and a limited number of shape features may compromise the classifier performance. Another possible reason might be that with a very limited number of features (only three in this case), the grammars used were not generic enough to deal with masses that were not present in the training set, which led to a misclassification.

Furthermore, on average, the **no-shape-label model** achieved an inferior result when compared to the other two models. This fact might be used to show the importance of the labels indicating whether the mass is circumscribed or spiculated when creating the hierarchical model in the classification process.

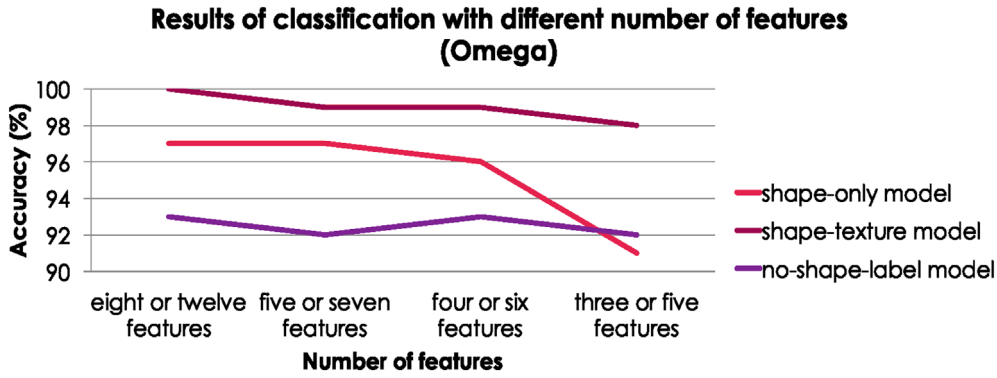


FIGURE 7 Accuracies achieved in the process of mass classification considering the three core hierarchical models and different number of features. The **shape-only model** used eight, five, four, and three features. The **shape-texture model** and the **no-shape-label model** used 12, 7, 6, and 5 features [Color figure can be viewed at wileyonlinelibrary.com]

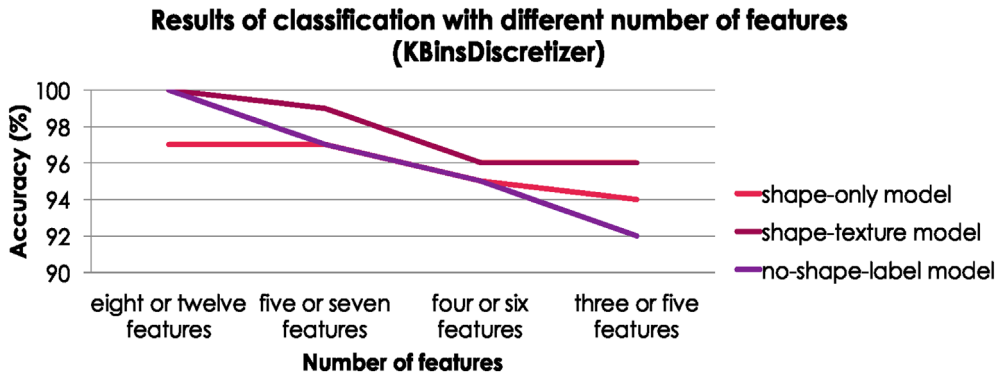


FIGURE 8 Accuracies achieved in the process of mass classification considering the three core hierarchical models and different number of features. The **shape-only model** used eight, five, four, and three features. The **shape-texture model** and the **no-shape-label model** used 12, 7, 6, and 5 features [Color figure can be viewed at wileyonlinelibrary.com]

Figure 9 shows the highest accuracies achieved by each core hierarchical model when all the features were used. We can observe the importance of the $H_{\min}$, as the higher this parameter is, the lower the accuracy obtained. All grammar-based models showed a good performance when $H_{\min} = 2$, but the accuracies decrease by around 10% when the $H_{\min} = 5$.

In Figure 10 we can see the highest accuracies when the KBinsDiscretizer algorithm is used in the feature discretization process and all the features are used to build the models. It can be noticed that as the parameter *n_bins* increases the accuracy of the model also increases. This behavior is the same presented when the Omega algorithm is used, but in that case the number of bins is influenced by the parameter $H_{\min}$.

The discretization process proved to be crucial to the performance of the proposed method. The $H_{\min}$ is an input parameter for Omega algorithm and restricts the minimum number of elements that each bin must have. In general, when the $H_{\min}$ value is high, less bins are obtained during the discretization process. Thus, the higher the $H_{\min}$ parameter, the higher the inconsistencies generated during this process. Considering this fact, it is important to keep this value as

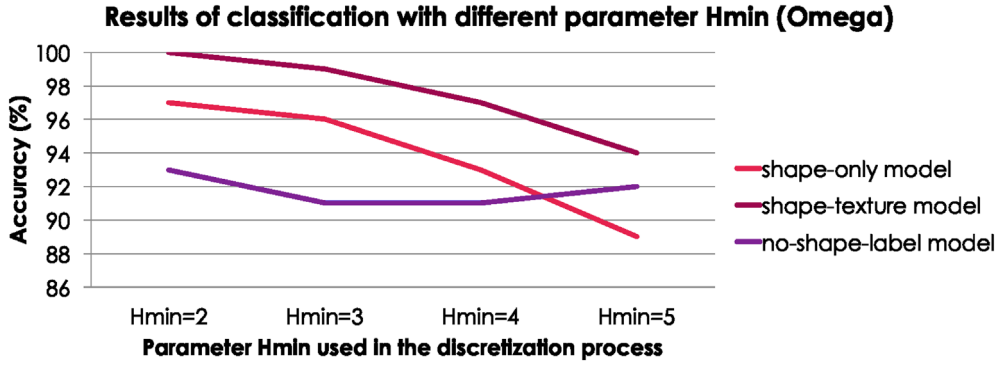


FIGURE 9 Accuracies achieved in the process of mass classification considering the three core hierarchical models, all (eight shape and four texture/gradient) features and different values of parameter $H_{\min}$ [Color figure can be viewed at wileyonlinelibrary.com]

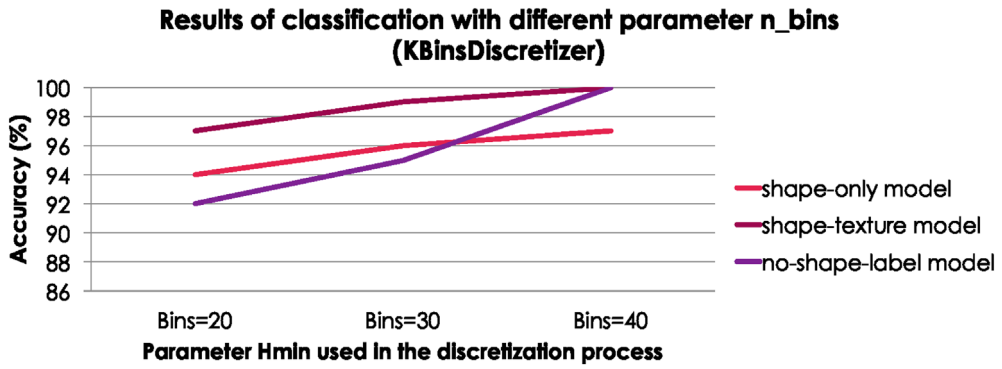


FIGURE 10 Accuracies achieved in the process of mass classification considering the three core hierarchical models, all (eight shape and four texture/gradient) features and different values of parameter *n_bins* [Color figure can be viewed at wileyonlinelibrary.com]

low as possible, even when only a small reduction in the number of bins is achieved.[14] The same behavior is also valid for the KBinsDiscretizer algorithm, that is, it is important to choose a proper value for the parameter *n_bins*, since the classification accuracies tend to be better for higher values of this parameter. In addition, the more inconsistencies in each bin, the worse the accuracies achieved by the models as shown in Figures 9 and 10.

Analyzing Figures 7, 8, 9, and 10 we can see that the **shape-texture model** achieved the best performance when the number of features and the number of bins vary. In general, the **shape-only model** has a better or the same performance than the **no-shape-label model** when the six or more features are used. However, when the number of features decrease, these two models tend to present a similar result.

Furthermore, the grammar-based models were compared with some of the most used classifiers found in the literature: ANN, SVM, KNN, and RF. The highest accuracies achieved by the grammar-based models have surpassed the highest accuracies achieved by the other classifiers by almost 10% in the experiments performed. However, these results do not imply that the traditional approaches should not continue being used to handle the mass classification problem. In fact, these results show that the grammar-based models had superior results in the classification task considering the images used and the features extracted. Moreover, these results show that syntactic approaches could be helpful and more explored to solve the mass classification problem. Figure 11 shows a comparison of the best results achieved by grammar-based models and by the traditional classifiers.

A limitation of using grammars, especially when compared to deep learning approaches, is the need to extract previously handcrafted features. To perform this extraction, a good segmentation method is required or the masses need to have their boundaries manually drawn by expert radiologists. The images used in this work had their boundaries manually drawn by experts, but images with boundaries drawn automatically by a segmentation algorithm could also have been used. Another limitation is that a discretization process is mandatory and, if the algorithm and the parameters chosen are not ideal, the classifiers might not perform as good as other techniques.

An advantage of hierarchical models and grammars is their concise representation and interpretability. In general, engineers and radiologists can easily understand the grammatical models, unlike ANNs, for instance. In this study, stochastic context-free grammars were used, but there are different grammatical approaches that a researcher can choose to deal with different

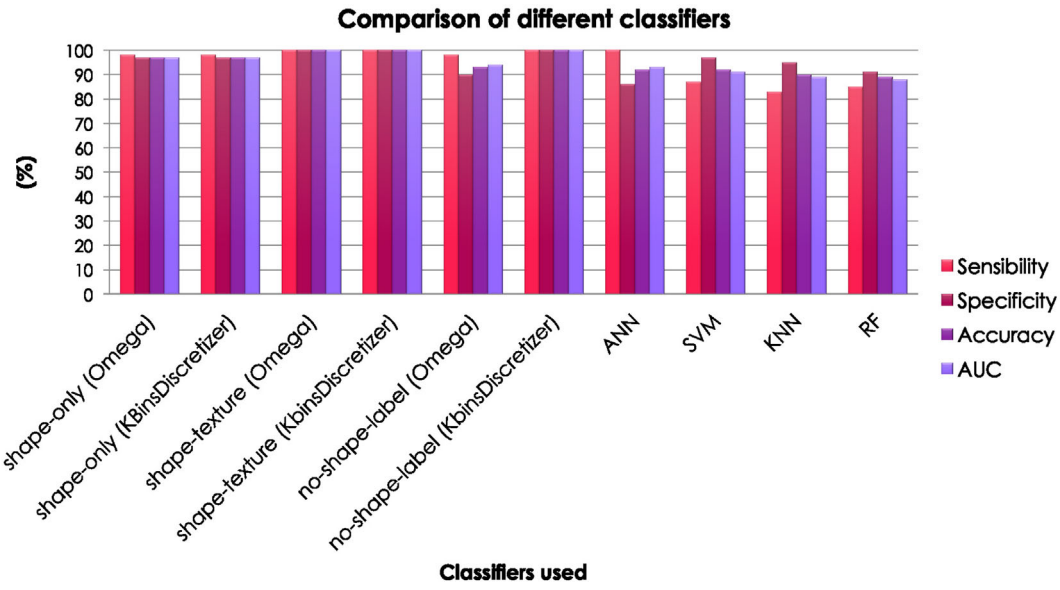


FIGURE 11 Highest accuracies achieved in the process of mass classification considering the three core hierarchical (shape-only, shape-texture and no-shape-label) models and the most used classifiers to discriminate masses [Color figure can be viewed at wileyonlinelibrary.com]

representations in order to solve this problem. Another advantage of using the proposed model, specially when compared to deep learning methods, is that the proposed model could learn the pattern of benign and malignant masses using a small number of images. In a previous work,[13] we used 57 and 54 images from two datasets in separate. In this work a combination of these images was used. For a deep learning approach it is necessary a larger dataset due to the complexity of the model and the number of features and weights to learn.

The proposed approach is formed by several intermediate steps, for example, feature discretization, hierarchical model representation, creation of stochastic grammars, and Bayesian classifier that could appear cumbersome, but other approaches also have intermediate steps. For instance, considering traditional ANNs it is generally required a standardization of features, the definition of the number of neurons and hidden layers, choose the right hyperparameters and learning the weights of the ANN before performing the classification. Thus, using syntactic approaches do not lead to a more complex process for the classifier learning.

8 | CONCLUSIONS

A syntactic approach to classify masses found in digital mammograms was presented in the present study. The tests were performed using 111 images from two different datasets, and the results show that the proposed syntactic approach can be helpful to classify masses as benign or malignant with superior results to the ones presented in recent papers.

Three core hierarchical models were created and the highest accuracies ranged from 0.98 to 1 when all features were used. A good discretization process proved to be mandatory, since depending on the parameters chosen, the accuracies can decrease by around 10%. Furthermore, the Gini importance of each feature was calculated and used to perform a feature selection.

The approaches demonstrated to be robust when the number of features decreased due to the feature selection implemented. The core model that used only shape features had a major decrease in its performance when fewer features were used. The other two core models proved to be more robust.

To compare the results obtained by the grammar-based models with non-hierarchical approaches, we implemented some of the most used machine learning techniques to handle the problem of mass classification (ANN, SVM, KNN, and RF). The grammar-based models showed better performance, in terms of accuracy, when compared with the more traditional models. In fact, the worst accuracies achieved by the grammar-based models were similar to the highest accuracies achieved by the traditional models (around 90%). In addition, the highest accuracies of the grammar-based models were almost 10% superior to the highest accuracies of the traditional models.

For future work, we are investigating how grammars can be used to generate synthetic masses that are similar to the real ones. These new generated masses could create a database of synthetic images that could be used by other researchers in projects of classification or segmentation of masses, as well as for training new radiologists.

Finally, we would like to mention how important this research area is to society as a whole. Breast cancer is one of the deadliest cancers that affects mainly women and deserves attention from research institutes.

ACKNOWLEDGMENTS

Brazilian National Council of Scientific and Technological Development (CNPq) (grant #309030/2019-6); CNPq and São Paulo Research Foundation (FAPESP): National Institute of Science and Technology – Medicine Assisted by Scientific Computing (INCT-MACC) – grant #157535/2017-7; FAPESP grant #2011/50761-2 and NAP eScience - PRP - USP.

We would also like to thank the authors of papers 19-23 for having provided the datasets used in this study.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

F.L.S.N and A.M.-L. designed and coordinated this research, reviewed and contributed with the discussions of this paper. R.W.D.P. performed the experiments and drafted the manuscript. The authors reviewed this document and approved the final manuscript.

ORCID

Ricardo Wandré Dias Pedro https://orcid.org/0000-0002-3728-4540

REFERENCES

1. Li H, Meng X, Wang T, Tang Y, Yin Y. Breast masses in mammography classification with local contour features. *BioMed Eng OnLine*. 2017;16:44. https://doi.org/10.1186/s12938-017-0332-0.
2. Mohanty AK, Senapati MR, Beberta S, Lenka SK. Texture-based features for classification of mammograms using decision tree. *Neural Comput Appl*. 2013;23(3):1011-1017. https://doi.org/10.1007/s00521-012-1025-z.

3. Todd CA, Golshah N. Method for breast cancer classification based solely on morphological descriptors. In: Michael FJ, Milan S, eds. *Medical Imaging 2004: Image Processing*. International Society for Optics and PhotonicsSPIE; 2004:857-867.
4. Keleş A, Keleş A, Yavuz U. Extracting fuzzy rules for the diagnosis of breast cancer. *Turkish J Electr Eng Comput Sci*. 2013;21(5):1495-1503.
5. Meriem H, Merouani HF, Lakhdar L. The power laws: Zipf and inverse Zipf for automated segmentation and classification of masses within mammograms. *Evolg Syst*. 2015;6(3):209-227.
6. Li Y, Chen H, Rohde GK, Yao C, Cheng L. Texton analysis for mass classification in mammograms. *Pattern Recogn Lett*. 2015;52(C):87-93. https://doi.org/10.1016/j.patrec.2014.10.008.
7. Xie W, Li Y, Ma Y. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*. 2016;173:930-941.
8. Dhungel N, Carneiro G, Bradley AP. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Med Image Anal*. 2017;37:114-128.
9. Pedro RWD, Machado-Lima A, Nunes FLS. Is mass classification in mammograms a solved problem? -a critical review over the last 20 years. *Expert Syst Appl*. 2019;119:90-103. https://doi.org/10.1016/j.eswa.2018.10.032.
10. Pedro RWD, Nunes FLS, Machado-Lima A. Using grammars for pattern recognition in images: a systematic review. *ACM Comput. Surv*. 2013;46(2):26:1-26:34. https://doi.org/10.1145/2543581.2543593.
11. Chang J, Yu J, Han T, Chang H, Park E. A method for classifying medical images using transfer learning: a pilot study on histopathology of breast cancer. Paper presented at: Proceedings of the 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom); 2017:1-4.
12. Tahmasbi A, Saki F, Shokouhi SB. CWLA: A novel cognitive classifier for breast mass diagnosis. Paper presented at: Proceedings of the 2011 18th Iranian Conference of Biomedical Engineering (ICBME); 2011:255-259.
13. Pedro RWD, Machado-Lima A, Nunes FLS. A new syntactic approach for masses classification in digital mammograms. Paper presented at: Proceedings of the 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); 2019:385-390.
14. Ribeiro MX, Ferreira MRP, Traina C Jr. Traina AJM. data pre-processing: a new algorithm for feature selection and data discretization. *CSTST '08*. New York, NY: ACM; 2008:252-257.
15. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
16. Panchal R, Verma B. Neural classification of mass abnormalities with different types of features in digital mammography. *Int J Comput Intell Appl*. 2006;6(1):61-75.
17. Gupta S, Chyn PF, Markey MK. Breast cancer CADx based on BI-RAds descriptors from two mammographic views. *Med Phys*. 2006;33(6):1810-1817. https://doi.org/10.1118/1.2188080.
18. Chomsky N. On certain formal properties of grammars. *Inf Control*. 1959;2(2):137-167. https://doi.org/10.1016/S0019-9958(59)90362-6.
19. Mudigonda NR, Rangayyan RM, Desautels JEL. Concavity and convexity analysis of mammographic masses via an iterative boundary segmentation algorithm. Paper presented at: Proceedings of the Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.99TH8411), vol. 3; 1999:1489-1494.
20. Rangayyan RM, Mudigonda NR, Desautels JEL. Boundary modelling and shape analysis methods for classification of mammographic masses. *Med Biol Eng Comput*. 2000;38(5):487-496. https://doi.org/10.1007/BF02345742.
21. Nandi RJ, Nandi AK, Rangayyan RM, Scutt D. Classification of breast masses in mammograms using genetic programming and feature selection. *Med Biol Eng Comput*. 2006;44(8):683-694.
22. Rangayyan RM, Nguyen TM. Fractal analysis of contours of breast masses in mammograms. *J. Dig Imag*. 2007;20(3):223-237.
23. Mudigonda NR, Rangayyan RM, Desautels JEL. Gradient and texture analysis for the classification of mammography masses. *IEEE Trans Med Imag*. 2000;19(10):1032-1043.
24. Alberta cancer board Canada. Screen test: alberta program for the early detection of breast cancer; 2004. http://www.cancerboard.ab.ca/screentest2001/03
25. Suckling J, Parker J, Dance DR, et al. The mammographic image analysis society digital mammogram database. Paper presented at: Proceedings of the 2nd International Workshop on Digital Mammography; 1994:375-378.

26. Chokri F, Farida MH. Mammographic mass classification according to Bi-RADS lexicon. *IET Comput Vis*. 2017;11(3):189-198. https://doi.org/10.1049/iet-cvi.2016.0244.
27. University of South Florida. Digital database for screening mammography; 2004. http://marathon.csee.usf.edu/Mammography/Database.html.
28. Jiao Z, Gao X, Wang Y, Li J. A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognit*. 2017;75:292-301. https://doi.org/10.1016/j.patcog.2017.07.008.
29. Kanadam KP, Chereddy SR. Mammogram classification using sparse-ROI: a novel representation to arbitrary shaped masses. *Expert Syst. Appl.* 2016;57:204-213.
30. Muramatsu C, Hara T, Endo T, Fujita H. Breast mass classification on mammograms using radial local ternary patterns. *Comp Bio Med*. 2016;72:43-53.
31. Menze BH, Kelm BM, Masuch E, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform*. 2009;10(1):213. https://doi.org/10.1186/1471-2105-10-213.
32. Liu X, Tang J. Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method. *IEEE Syst J*. 2014;8(3):910-920.
33. Zhu SC, Mumford D. A stochastic grammar of images. *Found Trends Comput Graph Vis*. 2006;2(4):259-362. https://doi.org/10.1561/0600000018.
34. Fu KS. *Syntactic Pattern Recognition and Applications*. Englewood Cliffs, NJ: Prentice-Hall, Inc.; 1982.
35. Earley J. An efficient context-free parsing algorithm. *Commun ACM*. 1970;13(2):94-102. https://doi.org/10.1145/362007.362035.
36. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Sattar A, Kang B-h, eds. *AI 2006: Advances in Artificial Intelligence*. Berlin, Heidelberg / Germany: Springer; 2006:1015-1021.

How to cite this article: Dias Pedro RW, Machado-Lima A, Nunes FLS. Towards an approach using grammars for automatic classification of masses in mammograms. *Computational Intelligence*. 2020;1–30. https://doi.org/10.1111/coin.12320

APPENDIX A. TABLES WITH ALL RESULTS

The results when KBinsDiscretizer was used in the feature discretization process are shown in Tables A5, A6, A7, and A8. The highest accuracy achieved was 0.97 using three, four, five, or all the shape features and $n\_bins = 40$. Moreover, the results using this algorithm showed a behavior similar to that observed using the Omega algorithm: model accuracy increase as the number of bins increases.

TABLE A1 Classification results using **shape-only**, **shape-texture** and **no-shape-label** models with Omega. Only the shape and gradient/texture features with the highest Gini importance were used to build the classifiers - (Shape features: Fractal dimension one-dimensional Ruler, Spiculation index, and Fractional concavity; Texture features: Contrast and Traditional acutance). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Three shape and two texture features - (Omega)						
Model	$H_{\min}$	$\zeta_{\max}$	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	**2**	**0.35**	**0.89** ± 0.25	**0.93 ± 0.19**	**0.91 ± 0.14**	**0.91 ± 0.15**
Shape-only	**2**	**0.40**	**0.89 ± 0.25**	**0.93 ± 0.19**	**0.91 ± 0.14**	**0.91 ± 0.15**
Shape-only	**2**	**0.45**	**0.89 ± 0.25**	**0.93 ± 0.19**	**0.91 ± 0.14**	**0.91 ± 0.15**
Shape-only	3	0.35	0.93 ± 0.17	0.94 ± 0.16	0.94 ± 0.11	0.94 ± 0.11

(continues)

TABLE A1 (Continued)

Three shape and two texture features - (Omega)						
Model	$H_{\min}$	$\zeta_{\max}$	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	3	0.40	0.93 ± 0.17	0.94 ± 0.16	0.94 ± 0.11	0.94 ± 0.11
Shape-only	3	0.45	0.93 ± 0.17	0.94 ± 0.16	0.94 ± 0.11	0.94 ± 0.11
Shape-only	4	0.35	0.91 ± 0.19	0.91 ± 0.20	0.91 ± 0.13	0.91 ± 0.12
Shape-only	4	0.40	0.91 ± 0.19	0.91 ± 0.20	0.91 ± 0.13	0.91 ± 0.12
Shape-only	4	0.45	0.91 ± 0.19	0.91 ± 0.20	0.91 ± 0.13	0.91 ± 0.12
Shape-only	5	0.35	0.89 ± 0.21	0.93 ± 0.17	0.91 ± 0.16	0.91 ± 0.17
Shape-only	5	0.40	0.89 ± 0.21	0.94 ± 0.16	0.92 ± 0.15	0.92 ± 0.16
Shape-only	5	0.45	0.89 ± 0.20	0.94 ± 0.16	0.92 ± 0.15	0.92 ± 0.16
Shape-texture	**2**	**0.35**	**0.96 ± 0.14**	**1**	**0.98 ± 0.06**	**0.98 ± 0.07**
Shape-texture	**2**	**0.40**	**0.96 ± 0.14**	**1**	**0.98 ± 0.06**	**0.98 ± 0.07**
Shape-texture	**2**	**0.45**	**0.96 ± 0.14**	**1**	**0.98 ± 0.06**	**0.98 ± 0.07**
Shape-texture	3	0.35	0.96 ± 0.14	0.96 ± 0.15	0.96 ± 0.10	0.96 ± 0.10
Shape-texture	3	0.40	0.96 ± 0.14	0.96 ± 0.15	0.96 ± 0.10	0.96 ± 0.10
Shape-texture	3	0.45	0.96 ± 0.14	0.96 ± 0.15	0.96 ± 0.10	0.96 ± 0.10
Shape-texture	4	0.35	0.96 ± 0.14	0.91 ± 0.24	0.93 ± 0.15	0.93 ± 0.13
Shape-texture	4	0.40	0.96 ± 0.14	0.91 ± 0.24	0.93 ± 0.15	0.93 ± 0.13
Shape-texture	4	0.45	0.96 ± 0.14	0.91 ± 0.24	0.93 ± 0.15	0.93 ± 0.13
Shape-texture	5	0.35	0.98 ± 0.10	0.97 ± 0.09	0.97 ± 0.07	0.97 ± 0.07
Shape-texture	5	0.40	0.93 ± 0.17	0.97 ± 0.09	0.96 ± 0.08	0.95 ± 0.09
Shape-texture	5	0.45	0.93 ± 0.17	0.97 ± 0.09	0.96 ± 0.08	0.95 ± 0.09
No-shape-label	**2**	**0.35**	**0.96 ± 0.14**	**0.90 ± 0.23**	**0.92 ± 0.15**	**0.93 ± 0.14**
No-shape-label	**2**	**0.40**	**0.96 ± 0.14**	**0.90 ± 0.23**	**0.92 ± 0.15**	**0.93 ± 0.14**
No-shape-label	**2**	**0.45**	**0.96 ± 0.14**	**0.90 ± 0.23**	**0.92 ± 0.15**	**0.93 ± 0.14**
No-shape-label	3	0.35	0.93 ± 0.17	0.90 ± 0.19	0.92 ± 0.13	0.92 ± 0.13
No-shape-label	3	0.40	0.93 ± 0.17	0.90 ± 0.19	0.92 ± 0.13	0.92 ± 0.13
No-shape-label	3	0.45	0.93 ± 0.17	0.90 ± 0.19	0.92 ± 0.13	0.92 ± 0.13
No-shape-label	4	0.35	0.91 ± 0.28	0.90 ± 0.23	0.91 ± 0.17	0.91 ± 0.17
No-shape-label	4	0.40	0.91 ± 0.25	0.90 ± 0.23	0.91 ± 0.16	0.91 ± 0.15
No-shape-label	4	0.45	0.91 ± 0.24	0.90 ± 0.23	0.91 ± 0.16	0.91 ± 0.15
No-shape-label	5	0.35	0.96 ± 0.20	0.90 ± 0.20	0.93 ± 0.14	0.93 ± 0.14
No-shape-label	5	0.40	0.93 ± 0.22	0.89 ± 0.26	0.91 ± 0.17	0.91 ± 0.16
No-shape-label	5	0.45	0.93 ± 0.22	0.89 ± 0.26	0.91 ± 0.17	0.91 ± 0.16

TABLE A2 Classification results using **shape-only**, **shape-texture**, and **no-shape-label** models with Omega. Only the shape and gradient/texture features with the highest Gini importance were used to build the classifiers - (Shape features: Fractal dimension one-dimensional Ruler, Spiculation index, Fractional concavity and Fourier factor; Texture features: Contrast and Traditional acutance). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Four shape and two texture features - (Omega)						
Model	H_{min}	ζ_{max}	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	**2**	**0.35**	**0.96 ± 0.14**	**0.97 ± 0.13**	**0.96 ± 0.10**	**0.96 ± 0.09**
Shape-only	**2**	**0.40**	**0.96 ± 0.14**	**0.97 ± 0.13**	**0.96 ± 0.10**	**0.96 ± 0.09**
Shape-only	**2**	**0.45**	**0.96 ± 0.14**	**0.97 ± 0.13**	**0.96 ± 0.10**	**0.96 ± 0.09**
Shape-only	3	0.35	0.96 ± 0.14	0.97 ± 0.09	0.96 ± 0.07	0.96 ± 0.08
Shape-only	3	0.40	0.96 ± 0.14	0.97 ± 0.09	0.96 ± 0.07	0.96 ± 0.08
Shape-only	3	0.45	0.96 ± 0.14	0.97 ± 0.09	0.96 ± 0.07	0.96 ± 0.08
Shape-only	4	0.35	0.96 ± 0.14	0.91 ± 0.20	0.93 ± 0.13	0.93 ± 0.11
Shape-only	4	0.40	0.96 ± 0.14	0.91 ± 0.20	0.93 ± 0.13	0.93 ± 0.11
Shape-only	4	0.45	0.96 ± 0.14	0.91 ± 0.20	0.93 ± 0.13	0.93 ± 0.11
Shape-only	5	0.35	0.89 ± 0.21	0.94 ± 0.16	0.92 ± 0.15	0.92 ± 0.16
Shape-only	5	0.40	0.89 ± 0.21	0.94 ± 0.16	0.92 ± 0.15	0.92 ± 0.16
Shape-only	5	0.45	0.89 ± 0.21	0.94 ± 0.16	0.92 ± 0.15	0.92 ± 0.16
Shape-texture	**2**	**0.35**	**0.98 ± 0.10**	**1**	**0.99 ± 0.04**	**0.99 ± 0.05**
Shape-texture	**2**	**0.40**	**0.98 ± 0.10**	**1**	**0.99 ± 0.04**	**0.99 ± 0.05**
Shape-texture	**2**	**0.45**	**0.98 ± 0.10**	**1**	**0.99 ± 0.04**	**0.99 ± 0.05**
Shape-texture	3	0.35	0.96 ± 0.14	0.98 ± 0.07	0.97 ± 0.07	0.97 ± 0.08
Shape-texture	3	0.40	0.96 ± 0.14	0.98 ± 0.07	0.97 ± 0.07	0.97 ± 0.08
Shape-texture	3	0.45	0.96 ± 0.14	0.98 ± 0.07	0.97 ± 0.07	0.97 ± 0.08
Shape-texture	4	0.35	0.98 ± 0.10	0.93 ± 0.19	0.95 ± 0.12	0.95 ± 0.10
Shape-texture	4	0.40	0.98 ± 0.10	0.93 ± 0.19	0.95 ± 0.12	0.95 ± 0.10
Shape-texture	4	0.45	0.98 ± 0.10	0.93 ± 0.19	0.95 ± 0.12	0.95 ± 0.10
Shape-texture	5	0.35	0.93 ± 0.17	0.97 ± 0.09	0.96 ± 0.08	0.95 ± 0.09
Shape-texture	5	0.40	0.93 ± 0.17	0.97 ± 0.09	0.96 ± 0.08	0.95 ± 0.09
Shape-texture	5	0.45	0.94 ± 0.17	0.97 ± 0.09	0.95 ± 0.08	0.95 ± 0.09
No-shape-label	**2**	**0.35**	**0.98 ± 0.10**	**0.90 ± 0.25**	**0.93 ± 0.16**	**0.94 ± 0.14**
No-shape-label	**2**	**0.40**	**0.98 ± 0.10**	**0.90 ± 0.25**	**0.93 ± 0.16**	**0.94 ± 0.14**
No-shape-label	**2**	**0.45**	**0.98 ± 0.10**	**0.90 ± 0.25**	**0.93 ± 0.16**	**0.94 ± 0.14**
No-shape-label	3	0.35	0.96 ± 0.14	0.86 ± 0.20	0.90 ± 0.13	0.90 ± 0.12
No-shape-label	3	0.40	0.96 ± 0.14	0.86 ± 0.20	0.90 ± 0.13	0.90 ± 0.12
No-shape-label	3	0.45	0.96 ± 0.14	0.86 ± 0.20	0.90 ± 0.13	0.90 ± 0.12
No-shape-label	4	0.35	0.93 ± 0.22	0.90 ± 0.23	0.92 ± 0.15	0.92 ± 0.15
No-shape-label	4	0.40	0.93 ± 0.22	0.90 ± 0.23	0.92 ± 0.15	0.92 ± 0.15
No-shape-label	4	0.45	0.93 ± 0.22	0.90 ± 0.23	0.92 ± 0.15	0.92 ± 0.15
No-shape-label	5	0.35	0.96 ± 0.20	0.90 ± 0.21	0.93 ± 0.14	0.93 ± 0.14
No-shape-label	5	0.40	0.96 ± 0.20	0.90 ± 0.21	0.93 ± 0.14	0.93 ± 0.14
No-shape-label	5	0.45	0.96 ± 0.20	0.90 ± 0.21	0.93 ± 0.14	0.93 ± 0.14

TABLE A3 Classification results using **shape-only**, **shape-texture,** and **no-shape-label** models with Omega. Only the shape and gradient/texture features with the highest Gini importance were used to build the classifiers - (Shape features: Fractal dimension one-dimensional Ruler, Fractal dimension two-dimensional Ruler, Spiculation index, Fractional concavity and Fourier factor; Texture features: Contrast, and Traditional acutance). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Five shape and two texture features - (Omega)						
Model	H_{min}	ζ_{max}	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	**2**	**0.35**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-only	**2**	**0.40**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-only	**2**	**0.45**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-only	3	0.35	0.96 ± 0.14	0.97 ± 0.09	0.96 ± 0.07	0.96 ± 0.08
Shape-only	3	0.40	0.96 ± 0.14	0.97 ± 0.09	0.96 ± 0.07	0.96 ± 0.08
Shape-only	3	0.45	0.96 ± 0.14	0.97 ± 0.09	0.96 ± 0.07	0.96 ± 0.08
Shape-only	4	0.35	0.93 ± 0.17	0.91 ± 0.20	0.92 ± 0.13	0.92 ± 0.12
Shape-only	4	0.40	0.93 ± 0.17	0.91 ± 0.20	0.92 ± 0.13	0.92 ± 0.12
Shape-only	4	0.45	0.93 ± 0.17	0.91 ± 0.20	0.92 ± 0.13	0.92 ± 0.12
Shape-only	5	0.35	0.87 ± 0.22	0.94 ± 0.16	0.91 ± 0.15	0.90 ± 0.16
Shape-only	5	0.40	0.87 ± 0.22	0.94 ± 0.16	0.91 ± 0.15	0.90 ± 0.16
Shape-only	5	0.45	0.87 ± 0.22	0.94 ± 0.16	0.91 ± 0.15	0.90 ± 0.16
Shape-texture	**2**	**0.35**	**0.98 ± 0.10**	**1**	**0.99 ± 0.04**	**0.99 ± 0.05**
Shape-texture	**2**	**0.40**	**0.98 ± 0.10**	**1**	**0.99 ± 0.04**	**0.99 ± 0.05**
Shape-texture	**2**	**0.45**	**0.98 ± 0.10**	**1**	**0.99 ± 0.04**	**0.99 ± 0.05**
Shape-texture	3	0.35	0.98 ± 0.10	0.98 ± 0.06	0.98 ± 0.06	0.98 ± 0.06
Shape-texture	3	0.40	0.98 ± 0.10	0.98 ± 0.06	0.98 ± 0.06	0.98 ± 0.06
Shape-texture	3	0.45	0.98 ± 0.10	0.98 ± 0.06	0.98 ± 0.06	0.98 ± 0.06
Shape-texture	4	0.35	0.98 ± 0.10	0.93 ± 0.19	0.95 ± 0.12	0.95 ± 0.10
Shape-texture	4	0.40	0.98 ± 0.10	0.93 ± 0.19	0.95 ± 0.12	0.95 ± 0.10
Shape-texture	4	0.45	0.98 ± 0.10	0.93 ± 0.19	0.95 ± 0.12	0.95 ± 0.10
Shape-texture	5	0.35	0.91 ± 0.19	0.97 ± 0.09	0.95 ± 0.10	0.94 ± 0.12
Shape-texture	5	0.40	0.91 ± 0.19	0.97 ± 0.09	0.95 ± 0.10	0.94 ± 0.12
Shape-texture	5	0.45	0.91 ± 0.19	0.97 ± 0.09	0.95 ± 0.10	0.94 ± 0.12
No-shape-label	**2**	**0.35**	**0.96 ± 0.14**	**0.90 ± 0.25**	**0.92 ± 0.16**	**0.93 ± 0.15**
No-shape-label	**2**	**0.40**	**0.96 ± 0.14**	**0.90 ± 0.25**	**0.92 ± 0.16**	**0.93 ± 0.15**
No-shape-label	**2**	**0.45**	**0.96 ± 0.14**	**0.90 ± 0.25**	**0.92 ± 0.16**	**0.93 ± 0.15**
No-shape-label	3	0.35	0.93 ± 0.17	0.86 ± 0.20	0.89 ± 0.13	0.90 ± 0.13
No-shape-label	3	0.40	0.93 ± 0.17	0.86 ± 0.20	0.89 ± 0.13	0.90 ± 0.13
No-shape-label	3	0.45	0.93 ± 0.17	0.86 ± 0.20	0.89 ± 0.13	0.90 ± 0.13
No-shape-label	4	0.35	0.89 ± 0.29	0.80 ± 0.19	0.90 ± 0.14	0.89 ± 0.16
No-shape-label	4	0.40	0.87 ± 0.30	0.89 ± 0.24	0.88 ± 0.17	0.88 ± 0.17
No-shape-label	4	0.45	0.87 ± 0.30	0.89 ± 0.24	0.88 ± 0.17	0.88 ± 0.17
No-shape-label	5	0.35	0.91 ± 0.28	0.90 ± 0.21	0.91 ± 0.16	0.91 ± 0.16
No-shape-label	5	0.40	0.93 ± 0.22	0.90 ± 0.21	0.92 ± 0.14	0.92 ± 0.14
No-shape-label	5	0.45	0.93 ± 0.22	0.90 ± 0.21	0.92 ± 0.14	0.92 ± 0.14

TABLE A4 Classification results using **shape-only**, **shape-texture**, and **no-shape-label** models with Omega. All features available were used to build the classifiers (Shape features: Compactness, Spiculation index, Fractional concavity, Fourier factor, Fractal dimension one-dimensional Ruler, Fractal dimension two-dimensional Ruler, Fractal dimension one-dimensional Box and Fractal dimension two-dimensional Box; Texture: Contrast, Accutance, Traditional acutance and Coeficient of variation). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Eight shape and four texture features (Omega)						
Model	H_{min}	ζ_{max}	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	**2**	**0.35**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-only	**2**	**0.40**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-only	**2**	**0.45**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-only	3	0.35	0.93 ± 0.17	0.97 ± 0.09	0.96 ± 0.08	0.95 ± 0.09
Shape-only	3	0.40	0.93 ± 0.17	0.97 ± 0.09	0.96 ± 0.08	0.95 ± 0.09
Shape-only	3	0.45	0.93 ± 0.17	0.97 ± 0.09	0.96 ± 0.08	0.95 ± 0.09
Shape-only	4	0.35	0.91 ± 0.19	0.94 ± 0.16	0.93 ± 0.11	0.93 ± 0.11
Shape-only	4	0.40	0.91 ± 0.19	0.94 ± 0.16	0.93 ± 0.11	0.93 ± 0.11
Shape-only	4	0.45	0.91 ± 0.19	0.94 ± 0.16	0.93 ± 0.11	0.93 ± 0.11
Shape-only	5	0.35	0.85 ± 0.27	0.94 ± 0.16	0.90 ± 0.15	0.89 ± 0.17
Shape-only	5	0.40	0.83 ± 0.28	0.94 ± 0.16	0.89 ± 0.15	0.88 ± 0.17
Shape-only	5	0.45	0.83 ± 0.28	0.94 ± 0.16	0.89 ± 0.15	0.88 ± 0.17
Shape-texture	**2**	**0.35**	**1**	**1**	**1**	**1**
Shape-texture	**2**	**0.40**	**1**	**1**	**1**	**1**
Shape-texture	**2**	**0.45**	**1**	**1**	**1**	**1**
Shape-texture	3	0.35	0.98 ± 0.10	1	0.99 ± 0.04	0.99 ± 0.05
Shape-texture	3	0.40	0.98 ± 0.10	1	0.99 ± 0.04	0.99 ± 0.05
Shape-texture	3	0.45	0.98 ± 0.10	1	0.99 ± 0.04	0.99 ± 0.05
Shape-texture	4	0.35	0.98 ± 0.10	0.97 ± 0.13	0.97 ± 0.09	0.97 ± 0.08
Shape-texture	4	0.40	0.98 ± 0.10	0.97 ± 0.13	0.97 ± 0.09	0.97 ± 0.08
Shape-texture	4	0.45	0.98 ± 0.10	0.97 ± 0.13	0.97 ± 0.09	0.97 ± 0.08
Shape-texture	5	0.35	0.91 ± 0.19	0.97 ± 0.09	0.95 ± 0.09	0.94 ± 0.10
Shape-texture	5	0.40	0.89 ± 0.20	0.97 ± 0.09	0.94 ± 0.09	0.93 ± 0.10
Shape-texture	5	0.45	0.89 ± 0.20	0.97 ± 0.09	0.94 ± 0.09	0.93 ± 0.10
No-shape-label	**2**	**0.35**	**0.98 ± 0.10**	**0.90 ± 0.21**	**0.93 ± 0.14**	**0.94 ± 0.12**
No-shape-label	**2**	**0.40**	**0.98 ± 0.10**	**0.90 ± 0.21**	**0.93 ± 0.14**	**0.94 ± 0.12**
No-shape-label	**2**	**0.45**	**0.98 ± 0.10**	**0.90 ± 0.21**	**0.93 ± 0.14**	**0.94 ± 0.12**
No-shape-label	3	0.35	0.93 ± 0.17	0.88 ± 0.21	0.90 ± 0.14	0.91 ± 0.13
No-shape-label	3	0.40	0.93 ± 0.17	0.90 ± 0.18	0.91 ± 0.13	0.92 ± 0.12
No-shape-label	3	0.45	0.93 ± 0.17	0.90 ± 0.18	0.91 ± 0.13	0.92 ± 0.12
No-shape-label	4	0.35	0.91 ± 0.24	0.89 ± 0.19	0.90 ± 0.15	0.90 ± 0.16
No-shape-label	4	0.40	0.91 ± 0.24	0.90 ± 0.19	0.91 ± 0.16	0.91 ± 0.16
No-shape-label	4	0.45	0.91 ± 0.24	0.90 ± 0.19	0.91 ± 0.16	0.91 ± 0.16
No-shape-label	5	0.35	0.91 ± 0.28	0.90 ± 0.19	0.91 ± 0.14	0.91 ± 0.16
No-shape-label	5	0.40	0.91 ± 0.28	0.92 ± 0.18	0.92 ± 0.14	0.92 ± 0.16
No-shape-label	5	0.45	0.91 ± 0.28	0.92 ± 0.18	0.92 ± 0.14	0.92 ± 0.16

TABLE A5 Classification results using **shape-only**, **shape-texture**, and **no-shape-label** models with KBinsDiscretizer. Only the shape and gradient/texture features with the highest Gini importance were used to build the classifiers - (Shape features: Fractal dimension one-dimensional Ruler, Spiculation index and Fractional concavity; Texture features: Contrast and Traditional acutance). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Three shape and two texture features (KBinsDiscretizer)					
Model	**Bins**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	20	0.91 ± 0.19	0.91 ± 0.20	0.91 ± 0.15	0.91 ± 0.15
Shape-only	30	0.93 ± 0.17	0.91 ± 0.24	0.92 ± 0.15	0.92 ± 0.14
Shape-only	**40**	**0.96 ± 0.14**	**0.93 ± 0.24**	**0.94 ± 0.15**	**0.94 ± 0.13**
Shape-texture	20	0.93 ± 0.16	0.93 ± 0.22	0.93 ± 0.15	0.93 ± 0.14
Shape-texture	30	0.96 ± 0.14	0.94 ± 0.18	0.95 ± 0.12	0.95 ± 0.11
Shape-texture	**40**	**0.98 ± 0.10**	**0.96 ± 0.20**	**0.96 ± 0.13**	**0.97 ± 0.11**
No-shape-label	20	0.87 ± 0.26	0.90 ± 0.23	0.89 ± 0.17	0.89 ± 0.17
No-shape-label	30	0.89 ± 0.25	0.90 ± 0.21	0.90 ± 0.15	0.90 ± 0.15
No-shape-label	**40**	**0.91 ± 0.24**	**0.93 ± 0.22**	**0.92 ± 0.16**	**0.92 ± 0.15**

TABLE A6 Classification results using **shape-only**, **shape-texture**, and **no-shape-label** models with KBinsDiscretizer. Only the shape and gradient/texture features with the highest Gini importance were used to build the classifiers - (Shape features: Fractal dimension one-dimensional Ruler, Spiculation index, Fractional concavity and Fourier factor; Texture features: Contrast and Traditional acutance). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Four shape and two texture features (KBinsDiscretizer)					
Model	**Bins**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	20	0.91 ± 0.19	0.93 ± 0.19	0.92 ± 0.15	0.92 ± 0.15
Shape-only	30	0.95 ± 0.14	0.94 ± 0.21	0.95 ± 0.13	0.95 ± 0.12
Shape-only	**40**	**0.96 ± 0.14**	**0.94 ± 0.19**	**0.95 ± 0.12**	**0.95 ± 0.11**
Shape-texture	20	0.89 ± 0.20	0.94 ± 0.21	0.92 ± 0.18	0.92 ± 0.18
Shape-texture	30	0.95 ± 0.14	0.97 ± 0.13	0.96 ± 0.10	0.96 ± 0.09
Shape-texture	**40**	**0.98 ± 0.10**	**0.96 ± 0.20**	**0.96 ± 0.13**	**0.97 ± 0.11**
No-shape-label	20	0.89 ± 0.25	0.89 ± 0.24	0.89 ± 0.19	0.89 ± 0.19
No-shape-label	30	0.91 ± 0.24	0.90 ± 0.21	0.90 ± 0.15	0.91 ± 0.15
No-shape-label	**40**	**0.98 ± 0.10**	**0.93 ± 0.22**	**0.95 ± 0.13**	**0.96 ± 0.12**

TABLE A7 Classification results using **shape-only**, **shape-texture**, and **no-shape-label** models with KBinsDiscretizer. Only the shape and gradient/texture features with the highest Gini importance were used to build the classifiers (Shape features: Fractal dimension one-dimensional Ruler, Fractal dimension two-dimensional Ruler Spiculation index, Fractional concavity and Fourier factor; Texture features: Contrast and Traditional acutance). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Five shape and two texture features (KBinsDiscretizer)					
Model	**Bins**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	20	0.89 ± 0.20	0.94 ± 0.16	0.92 ± 0.11	0.92 ± 0.12
Shape-only	30	0.96 ± 0.14	0.96 ± 0.15	0.96 ± 0.10	0.96 ± 0.10
Shape-only	**40**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-texture	20	0.91 ± 0.19	0.97 ± 0.09	0.94 ± 0.10	0.94 ± 0.12
Shape-texture	30	0.96 ± 0.14	0.98 ± 0.07	0.97 ± 0.07	0.97 ± 0.08
Shape-texture	**40**	**0.98 ± 0.10**	**1**	**0.99 ± 0.04**	**0.99 ± 0.05**
No-shape-label	20	0.89 ± 0.25	0.90 ± 0.16	0.90 ± 0.13	0.90 ± 0.14
No-shape-label	30	0.91 ± 0.24	0.95 ± 0.13	0.93 ± 0.12	0.93 ± 0.13
No-shape-label	**40**	**0.98 ± 0.10**	**0.96 ± 0.12**	**0.97 ± 0.07**	**0.97 ± 0.07**

TABLE A8 Classification results using **shape-only**, **shape-texture**, and **no-shape-label** models with KBinsDiscretizer. All features available were used to build the classifiers (Shape features: Compactness, Spiculation index, Fractional concavity, Fourier factor, Fractal dimension one-dimensional Ruler, Fractal dimension two-dimensional Ruler, Fractal dimension one-dimensional Box and Fractal dimension two-dimensional Box; Texture: Contrast, Accutance, Traditional acutance, and Coefficient of variation). In this table the highest accuracies and area under the ROC curve (AUC) achieved by each classifier are shown in bold

Eight shape and four texture features (KBinsDiscretizer)					
Model	**Bins**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
Shape-only	20	0.93 ± 0.17	0.94 ± 0.16	0.94 ± 0.11	0.94 ± 0.11
Shape-only	30	0.98 ± 0.10	0.96 ± 0.15	0.96 ± 0.10	0.97 ± 0.09
Shape-only	**40**	**0.98 ± 0.10**	**0.97 ± 0.13**	**0.97 ± 0.09**	**0.97 ± 0.08**
Shape-texture	20	0.98 ± 0.10	0.97 ± 0.09	0.97 ± 0.07	0.97 ± 0.07
Shape-texture	30	1	0.98 ± 0.07	0.99 ± 0.04	0.99 ± 0.03
Shape-texture	**40**	**1**	**1**	**1**	**1**
No-shape-label	20	0.93 ± 0.17	0.90 ± 0.16	0.92 ± 0.10	0.92 ± 0.10
No-shape-label	30	0.96 ± 0.14	0.95 ± 0.13	0.95 ± 0.09	0.95 ± 0.09
No-shape-label	**40**	**1**	**1**	**1**	**1**

TABLE A9 The best results achieved by each of the most used classifiers. Features: Compactness (CC), Spiculation Index (SI), Fractional Concavity (FC), Fourier Factor (FF), Fractal dimension two-dimensional box counting (2B), Fractal dimension one-dimensional box counting (1B), Fractal dimension two-dimensional ruler (2R), Fractal dimension one-dimensional ruler (1R), Contrast (CO), Acutance (AC), Traditional acutance (TA) and Coefficient of Variation (CV)

Best results using ANN, SVM, KNN, and RF					
Classifier	**Feature combination**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
ANN	CC, FC, 2B, 1B, 2R, AC	1	0.86 ± 0.20	0.92 ± 0.11	0.93 ± 0.10
SVM	SI, FC, FF, CO, CV	0.87 ± 0.26	0.97 ± 0.11	0.92 ± 0.13	0.91 ± 0.14
KNN	SI, 2R, CO, CV	0.83 ± 0.24	0.95 ± 0.13	0.90 ± 0.11	0.89 ± 0.12
RF	2R, 1R, CO, AC	0.85 ± 0.23	0.91 ± 0.14	0.89 ± 0.13	0.88 ± 0.14

Abbreviations: ANN, artificial neural networks; KNN, k-nearest neighborhood; RF, random forest; SVM support vector machines.

TABLE A10 The worst results achieved by each of the most used classifiers. Features: Compactness (CC), Spiculation Index (SI), Fractional Concavity (FC), Fourier Factor (FF), Fractal dimension two-dimensional box counting (2B), Fractal dimension one-dimensional box counting (1B), Fractal dimension two-dimensional ruler (2R), Fractal dimension one-dimensional ruler (1R), Contrast (CO), Acutance (AC), Traditional acutance (TA) and Coefficient of Variation (CV)

Worst results using ANN, SVM, KNN, and RF					
Classifier	**Feature combination**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
ANN	1R, CO, AC	0	1	0.58 ± 0.06	0.5
SVM	CO, AC, CV	0.30 ± 0.31	0.57 ± 0.39	0.45 ± 0.16	0.43 ± 0.14
KNN	TA	0.22 ± 0.25	0.54 ± 0.29	0.40 ± 0.20	0.38 ± 0.20
RF	CO	0.41 ± 0.35	0.55 ± 0.32	0.49 ± 0.19	0.48 ± 0.20

Abbreviations: ANN, artificial neural networks; KNN, k-nearest neighborhood; RF, random forest; SVM support vector machines.