Using Support Vector Machine and Features Selection on Classification of Early Lung Nodules

Lucas L. Lima[1], **Thales Vieira**[2], **Evandro Barros Costa**[2], **Paulo M. Azevedo-Marques**[1], **Marcelo C. Oliveira**[2]

[1] Programa de Pós-graduação Interunidades em Bioengenharia, Universidade de São Paulo – Ribeirão Preto, SP – Brasil

[2]Instituto de Computação, Universidade Federal de Alagoas (UFAL) Maceió, AL – Brasil

lucaslima7@usp.br, pmarques@fmrp.usp.br,

{thales,evandro,oliveiramc}@ic.ufal.br

Resumo. *O câncer de pulmão é o câncer que mais mata no mundo. No entanto, se o diagnóstico for feito no início da doença, as taxas de sobrevida em 1 ano são de aproximadamente 81-85%. Ferramentas de Auxílio ao Diagnóstico por Computador têm um grande potencial para auxiliar os especialistas na determinação da malignidade de um nódulo pulmonar. Neste trabalho, foram utilizados 4 grupos de atributos: Textura 3D, Nitidez da margem 3D, Forma 3D e Intensidade 3D; dois algoritmos de aprendizado de máquina: Support Vector Machine (SVM) e Multilayer Perceptron; e duas técnicas para selecionar os recursos mais relevantes: Relief e Algoritmo Genético Evolucionário (AGE). A classificação com SVM, Relief e AGE alcançou a melhor AUC de 0,856.*

Abstract. *Lung cancer the cancer that kills most in the world. However, if the diagnosis is made at the beginning of the disease, the 1-year survival rates are approximately 81-85%. Computer-Aided Diagnosis tools have a great potential to auxiliary the experts in determining the malignancy of a lung nodule. In this work, we used 4 groups of features: 3D Texture, 3D Margin Sharpness, 3D Shape, and 3D Intensity; two machine learning algorithms: Support Vector Machine (SVM) and Multilayer Perceptron; and two techniques to select the most relevant features: Relief and Evolutionary Genetic Algorithm (EGA). The classification with SVM, Relief, and EGA achieved the best AUC of 0.856.*

1. Introdução

Among all cancer-related deaths in the world, lung cancer is the leading cause, accounting for approximately 20% [Tammemagi and Lam 2014]. For the current year, 2020, INCA estimates 30,200 new cases of lung cancer in Brazil, with 17,760 men and 12,440 women [INCA 2020]. In the United States, lung cancer is among the most lethal cancers in both men and women [Siegel et al. 2017].

From the moment the diagnosis of lung cancer is confirmed, over half of the patients die within one year, and the 5-year survival rate is around 17.8% [Zappa and Mousa 2016]. The survival rate is related to the lung cancer stage at diagnosis; for example, a patient diagnosed with one-year in advanced disease (stage-IV

with metastasis) has a survival rate of approximately 15-19%. However, 1-year survival rates could increase to the 81-85% range when the disease is diagnosed at an early stage (stage-I) [Neal et al. 2019, Knight and et 2017]. Thus, considering that a nodule may be a manifestation of cancer, the early detection, and measurement of pulmonary nodules are crucial to increase the chances of survival of cancer patients, in part due to the larger range of feasible treatments [Wu and et. al. 2013, Revees and et al 2006].

The main manifestation of lung cancer is through the pulmonary nodule. Computed tomography (CT) exams have been widely used by radiologists to diagnose lung cancer because it provides high-resolution 3D images with high contrast able to show differences in tumors' size, shape (e.g., rounded or spiculated), and texture (e.g., calcification). However, the interpretations of a medical image by professionals have shown significant variation in numerous studies due to several aspects, for example: rush for the results, recognition of variations between readers based on perceptual errors, lack of training, or fatigue [Akgul and et. al. 2011]. Other factors that make the diagnosis a challenging task for the radiologists are the low contrast, especially in early stages (with up to 10mm in diameter size), where the nodules may be attached to complex structures of the lung (Figure 1).

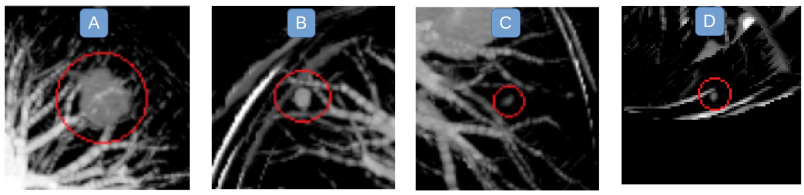

Figure 1. Examples of pulmonary nodules (highlighted in red). (A) 23.1mm connected to pleura; (B) 8.3mm connected to pleura; (C) 6.2mm isolated; (D) 5.7mm isolated [Alilou and et al 2014].

A way to aid radiologists in the process to diagnose small pulmonary nodules is to integrate the Computer-Aided Diagnosis (CADx) to imaging interpretation. CADx systems aim to help specialists improve diagnostic accuracy by acting as a second opinion for them through a computer-supplied suggestion without making the final diagnosis [Gillies et al. 2016]. Therefore, the adoption of CADx systems is appropriate and desirable to the process of diagnosis and interpretation of lung nodules on CT images [Chuquicusma and et al 2017, Ferreira-Junior et al. 2016].

When nodules are already segmented from computer tomography images of the lung, the first step in the CAD's systems is to extract features from the nodules. Some kinds of features to represent the nodules are: geometric, use of histograms, textures, shapes, margins, and densities [Shewaye and Mekonnen 2016, Choi and T. 2014, junior et al. 2016]. The next step is to classify the nodule into benign or malignant using classifiers such as Artificial Neural Networks, Logistic Regression, Support Vector Machines and, Deep Learning [Ciompi and et al 2017, Revees and et al 2006, Felix et al. 2016].

Some works faced the problem of classifying small lung nodules, but the AUC-ROC was not satisfactory [Felix et al. 2016, Reeves et al. 2015, Yan and et 2018]. So this problem, which is a few explored, is opened in the literature and needs to be further

investigated.

The objective of this work was to evaluate the accuracy of the classification of early pulmonary nodules, whose diameters are between 3-10 mm, in benign or malignant using 4 categories of attributes extracted from the nodules, which are: 3D Texture Features (TF), 3D Margin Sharpness Features (MSF), 3D Shape and 3D Intensity.

The remainder of this paper is organized as follows. First, we describe how was done the preparation of the database used and the methodology applied (Section 2). Next, we show and discuss the results of this work related to the state of art (Section 3. To finish, section 4 finishes this work.

2. Methodology

An overview of our method is shown in figure 2. First (A), is presented the preparation of the database of images from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC) (Section 2.1). In the pre-processing step (B), we measured the nodule size (Section 2.2), selected small nodules (Section 2.3), later the features were extracted from the filtered nodules (Section 2.4), and the most important features were selected (Section 2.5). Finally, the nodules were classified as benign or malignant in the last step (Section 2.6).

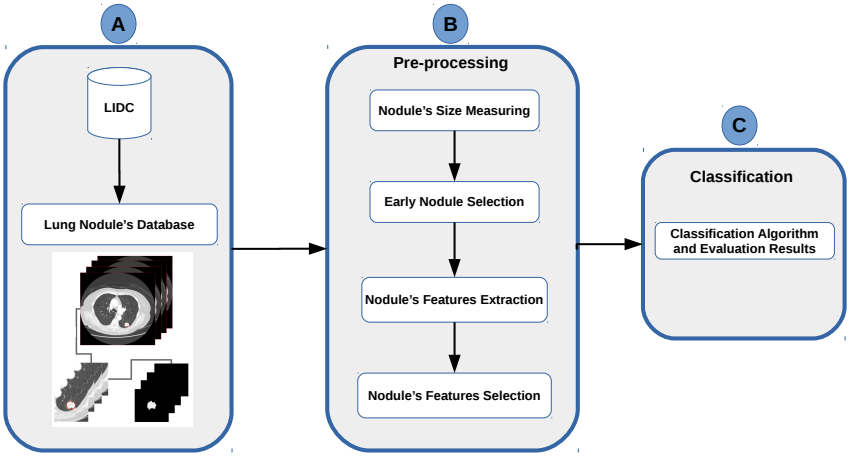

Figure 2. General schema of this work.

The steps of features selection and classification of the nodules were executed using the tool Rapidminer Studio [Lee and et al]. The tests were performed on a PC Intel Core i5, 3.10Hz CPU and 8GB RAM running GNU/Linux Ubuntu 16.04 LTS.

2.1. Lung Nodule's Database

We used the nodule images from the LIDC, which is a public database. It consists of CT scans for lung cancer with masked-up annotated lesions and classified by four experienced radiologists, including nodule outlines and subjective nodule characteristic ratings, in a procedure of image interpretation which required the specialists to read the CT scans and marking the lesions using a graphical interface [Armato and et al 2011]. To our knowledge, this database is the most important and used lung cancer image collection for CAD developers and researchers [junior et al. 2016].

The nodules identified by the radiologists were ranked according to subjective characteristics, and one of such characteristics was the malignancy on a 1-5 scale following the conditions: Malignancy 1: high probability of being benign; Malignancy 2: a moderate probability of being benign; Malignancy 3: indeterminate probability; Malignancy 4: a moderate probability of being malignant; Malignancy 5: high probability of being malignant.

In the LIDC there were four radiologist's marks, and we chose only one mark to avoid redundancies, using as the criterion the mark from the radiologist who identified the highest number of lesions in each exam. The pulmonary nodules with suspicious of malignancy 1 and 2 were considered to be benign, and pulmonary nodules with the probability of malignancy 4 and 5 were considered to be malignant.

2.2. Nodule Size Measuring

To perform the selection of the nodules according to their diameters, we initially estimate their sizes which were then attached to our database, as they were not given in the LIDC original database. The size of a nodule was estimated as a simple 2D measure of the greatest diameter, which can be performed along the x-axis and y-axis planes of the biggest diameter [Bartholmai and et al 2015]. These approximations consisted of calculating the Euclidean distance between the minimum and maximum coordinates in the respective *x* and *y* axes of all slices of a nodule. The biggest distance found was selected to be the diameter of the respective nodule.

2.3. Early Nodule Selection

After estimating the nodules diameters, we set thresholds to select early nodules by size. According to Bartholmai et. al. [Bartholmai and et al 2015], nodules smaller than 10mm still have chances to be malignant, and nodules bigger than 10mm have much more probability to be malignant. Note that the smallest nodule in our database has 3.27mm in diameter and that we are not interested in large nodules. Thus, we empirically set the thresholds to only select nodules whose diameters are between 3 mm to 10 mm.

To achieve fair classification, we balanced the number of benign and malignant nodules in 158 samples for each, a total of 316 samples. This was required because our database was composed of much more benign nodules compared to malignant nodules. This was expected due to the higher chances of early nodules to be benign [Bartholmai and et al 2015].

2.4. Nodule's Features Extraction

In this step, 71 features were extracted from the nodules selected. Such descriptors provided information represented by numeric values that are later concatenated in feature vectors. Four kinds of attributes were used: 3D Texture Features, 3D Margin Sharpness Features, 3D Intensity Features, and 3D Shape Features.

The 3D Texture Features (3DTF) extracted followed the proposal of Haralick et al. [Haralick et al. 1973], which are:

$$\text{Energy} = \sum_{i,j} C^2(i,j), \tag{1}$$

$$\text{Entropy} = -\sum_{i,j} C(i,j) log C(i,j), \tag{2}$$

$$\text{Inverse difference moment} = \sum_{i,j} \frac{C(i,j)}{1+(i-j)^2}, \tag{3}$$

$$\text{Shade} = \sum_{i,j} (i+j-\mu_x-\mu_y)^3 C(i,j), \tag{4}$$

$$\text{Inertia} = \sum_{i,j} (i-j)^2 C(i,j), \tag{5}$$

$$\text{Variance} = \sum_{i,j} (i-\mu)^2 C(i,j), \tag{6}$$

$$\text{Promenance} = \sum_{i,j} (i+j-\mu_x-\mu_y)^4 C(i,j), \tag{7}$$

$$\text{Correlation} = -\sum_{i,j} \frac{(i-\mu_x)(j-\mu_y)}{\sqrt{\sigma_x \sigma_y}} C(i,j), \tag{8}$$

$$\text{Homogeneity} = \sum_{i,j} \frac{C(i,j)}{1+|i-j|}, \tag{9}$$

where, $C(i,j)$ are elements $[i,j]$ of the co-occurrence matrix, μ_x and μ_y are averages, and σ_x and σ_y are standard deviations obtained from the equations below:

$$\mu_x = \sum_i i C_x(i), \tag{10}$$

$$\mu_y = \sum_j j C_y(j), \tag{11}$$

$$\sigma_x = \sum_i (i-\mu_x)^2 \cdot \sum_j C(i,j), \tag{12}$$

$$\sigma_y = \sum_j (j-\mu_y)^2 \cdot \sum_i C(i,j), \tag{13}$$

$$C_x(i) = \sum_j C(i,j), \tag{14}$$

$$C_y(j) = \sum_i C(i,j). \tag{15}$$

For each image, there will always be four co-occurrence matrices, each one corresponding to each direction (0°, 45°, 90° and 135°, counterclockwise). Thus, the calculus of the attributes applied to the co-occurrence matrices in the 4 orientations composes a TF vector with a 36-dimensional.

The 3D Margin Sharpness Features (3DMSF) extracted in this work were based on a statistical analysis, following a model proposed by Xu et al. [Xu and et al 2012]. The 12 attributes extracted were:

$$\text{Difference of two ends} = x_n - x_1, \tag{16}$$

$$\text{Sum of values} = \sum_{i=1}^{n} x_i, \tag{17}$$

$$\text{Sum of squares} = \sum_{i=1}^{n} x_i^2, \tag{18}$$

$$\text{Sum of logs} = \sum_{i=1}^{n} \log x_i, \tag{19}$$

$$\text{Arithmetic mean } (\mu) = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{20}$$

$$\text{Geometric mean} = \sqrt[n]{\prod_{i=1}^{n} x_i}, \tag{21}$$

$$\text{Population variance} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2, \tag{22}$$

$$\text{Sample variance } (\upsilon) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2, \tag{23}$$

$$\text{Standard deviation } (s) = \sqrt{\upsilon}, \tag{24}$$

$$\text{Kurtosis measure} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^4}{s^4}, \tag{25}$$

$$\text{Skewness measure} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^3}{s^3}, \tag{26}$$

$$\text{Second central measure} = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}{s^2}, \tag{27}$$

where, x is the intensity value of the pixel array of size n, x_1 is the intensity value of the pixel outside the nodule region, and x_n is the pixel intensity value in the region inside of the nodule. Thus, each nodule is associated with a 12-dimensional MSF vector.

The 3D Intensity Features (3DIF) were suggested by Dilger [Dilger and et al 2015]. The features are:

$$\text{Energy} = \sum_{i}^{n} x_i^2, \tag{28}$$

$$\text{Average intensity } (\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{29}$$

$$\text{Median intensity}, \tag{30}$$

$$\text{Minimum intensity } (I_m), \tag{31}$$

$$\text{Maximum intensity } (I_M), \tag{32}$$

$$\text{Entropy} = -\sum_{k=1}^{N} p(x_k) \log_2(p(x_k)), \tag{33}$$

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)s^4}, \tag{34}$$

$$\text{Skewness} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}, \tag{35}$$

$$\text{Absolute mean deviation} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|, \tag{36}$$

$$\text{Range} = |I_M - I_m|, \tag{37}$$

$$\text{Square root mean} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}, \tag{38}$$

$$\text{Standard deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{39}$$

$$\text{Uniformity} = \sum_{k=1}^{N} p(x_k)^2, \tag{40}$$

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{41}$$

where, *x* is the i-th intensity value of the image, *s* is the standard deviation of the intensity, *N* is the number of pixels in the region, and $p(x_k)$ is the probability of occurrence of the *kth* value of intensity in a set of *N* intensities. So, each nodule was associated with an IF vector with dimension 14.

The 3D Shape Features (3DSF) proposed by Aerts [Aerts and et al 2014] were adapted and implemented in this work:

$$\text{Compactness 1} = \frac{V}{\sqrt{\pi}A^{\frac{2}{3}}}, \tag{42}$$

$$\text{Compactness 2} = 36\pi\frac{V^2}{A^3}, \tag{43}$$

$$\text{Spherical disproportion} = \frac{A}{4\pi R^2}, \tag{44}$$

$$\text{Sphericity} = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}, \tag{45}$$

$$\text{Area } (A_1) = \sum^{i\in f} p_i, \tag{46}$$

$$\text{Surface area } (A_2) = 4\pi R^2, \tag{47}$$

$$\text{Surface-volume ratio} = \frac{A}{V}, \tag{48}$$

$$\text{Volume } (V) = (\sum^{f\in F}\sum^{i\in f} p_i)\cdot thicknessofF, \tag{49}$$

where, p_i represents the i-th pixel of the segmentation slice; f represents the set of pixels (p); F represents the total set of slices (f) in the nodule; $Fthickness$ is the size of the *voxel*; R represents the radius of the sphere with the same volume as the tumor,

defined as:

$$\text{Radius } (R) = \sqrt[3]{\frac{3V}{4\pi}} \tag{50}$$

Thus, each nodule was associated with a 9-dimensional F vector.

2.5. Nodule's Features Selection

To select the most relevant features for classification of small pulmonary nodules. The techniques we tried were the Evolutionary Genetic Algorithm (EGA) and Relief. We applied them on each group of attributes, and the integration of the four kinds of attributes extracted from the nodules.

EGA is based on genetic and evolutionary theory, wherein the chromosome bits represent if the feature is included or not [Chandrashekar and Sahin 2014]. The advantage of this method is the tendency to always select the fittest individuals. Relief has the objective to estimate the quality of features according to how well their values distinguish between the instances of the same and different classes, applying weights to each feature. Selecting the normalization in the relief's results, the features coming in rank between 0 and 1.

In our work, the population is formed by the features extracted from the nodules, so the EGA works on it generating a subset of the most relevant features in a binary way, where 1 indicates relevant features and 0 indicates irrelevant features. Individuals selected reproduction were chosen using tournament criteria. In the reproductive phase, chosen operators were: crossover and mutation, with probabilities of 50% and 5%, respectively. The crossover type applied was one-point. The population size was 40 and the maximum number of generations was 50.

2.6. Classification

The classifiers tried in this work were Support Vector Machines (SVM) with the Gaussian kernel and the Artificial Neural Network (ANN) technique called Multilayer Perceptron (MLP). The SVM is a supervised technique used for classification and regression analysis. The objective of the SVM is to find a hyperplane between data from two distinct classes possibly nonlinearly mapped to space where they are separable [Scholkopf and Smola 2001]. Some previous works have used SVM to detect and classify pulmonary nodules [Choi and Choi 2013, Madero Orozco and et al 2015].

Five sets of features were evaluated utilizing the SVM classifier: 1) 3DTF; 2) 3DMSF; 3) 3DSF; 4) 3DIF; 5) The integration of the four kinds of features (3DTF, 3DMSF, 3DSF and 3DIF). The total number of features used is this work was of 71 features (36 3DTF + 12 3DMSF + 14 3DSF + 9 3DIF). We performed an 10-fold cross validation to calibrate the SVM's hyperparameters.

We evaluated variations of the proposed method. We used the Area Under the ROC Curve (AUC) to evaluate the performance of: (1) Support Vector Machine (SVM) using Relief combined with EAG; (2) SVM using Relief only; (3) SVM using EAG only; and (4) Multilayer Perceptron (MLP) combined with EAG. The objective of the last combination (4) was to compare SVM with the best result obtained by Felix et al. [Felix et al. 2016], where they best result was using MLP. We also tried to evaluate the use of partial combinations of the features and all features together.

3. Results and Discussion

Table 1 shows results using SVM with and without EAG. Table 2 shows results using Relief with and without EAG. Table 3 shows results using MLP with EAG, which is similar to the approach of Felix et al. [Felix et al. 2016], but applied to the novel feature sets proposed in this work. The tables show classification results (mean $\pm$ standard deviation) over a 10-fold cross validation of the classifiers.

Table 1. Early pulmonary nodule classification using SVM with and without EAG.

Classifier	Group	AUC	Use of EAG
SVM	**3DTF**	0.816 +/- 0.082	Yes
		0.792 +/- 0.067	No
	3DMSF	0.749 +/- 0.092	Yes
		0.750 +/- 0.044	No
	3DSF	0.807 +/- 0.058	Yes
		0.783 +/- 0.099	No
	3DIF	0.817 +/- 0.072	Yes
		0.783 +/- 0.076	No
	All Features	**0.848 +/- 0.079**	Yes
		0.701 +/- 0.070	No

Table 2. Early pulmonary nodule classification using SVM with Relif and EAG.

Classifier	Group	AUC	
		Relief	**Relief with AG**
SVM	**3DTF**	0.751 ./- 0.065	0.780 +/- 0.081
	3DMSF	0.746 +/- 0.074	0.779 +/- 0.064
	3DSF	**0.807 +/- 0.070**	0.817 +/- 0.061
	3DIF	0.772 +/- 0.082	0.800 +/- 0.081
	All Features	0.777 +/- 0.069	**0.856 +/- 0.027**

In Table 1, the best result was achieved using all features together with the EAG, showing AUC of 0.848 (σ = 0.079), and the best result without EAG was achieved by the Texture Features, showing AUC of 0.792. We can conclude that using EAG is indeed relevant to select features that improve SVM classification. It is also worth mentioning that by using all features without feature selection, an AUC of only 0.701 is achieved. Another observation is that only for Margin Sharpness Features (MSF) the use of EAG is irrelevant. The number of features chosen by EAG in our best result (all features) was 26, showing a considerable reduction from the initial 71 features.

In Table 2, which shows results using Relief, the best AUC was achieved using all features with Relief and EAG, with an AUC of 0.856, the best result using only Relief was achieved using the Shape Features (SF) with AUC of 0.807 (σ = +/- 0.070). By comparing with the best result of Table 1, we conclude that the use of Relief only slightly improves the classification, and thus it is not considered relevant when used together with EAG.

Table 3. Early pulmonary nodule classification using MLP with EAG.

Classifier	Group	AUC
MLP with EAG	**ALL Features**	**0.836 +/- 0.071**

As shown in Table 3, the AUC of 0.836 (σ = 0.071) indicates that the SVM classifier using the Gaussian Kernel performed better than MLP and EAG in all features, when comparing to the results presented in Table 1. Finally, from our experiments we conclude that SVM with relief and EAG is the best combination to classify early pulmonary nodules.

We also compared in table 4 our best result with the best results of [Revees and et al 2006], [Felix et al. 2016] and [Dhara and et al 2016]. Felix et al. [Felix et al. 2016] obtained the best result (AUC of 0.820) using the MLP classifier with EAG, using 48 features (Texture Features with Margin Sharpness Features) and they used small pulmonary nodules exclusively, with diameter between 3-10mm. Reeves et al. [Revees and et al 2006] used 46 3D features such as geometry features, features of density distribution, surface curvature features and margin features, in a total of 326 nodules balanced between benign and malignant, with a diameter size between 5-14mm, and achieved an average of AUC of 0.708. Dhara et al. [Dhara and et al 2016] used 49 2D and 3D features, obtaining the best AUC of 0.950. Among the related works, the work of Felix et al. [Felix et al. 2016] is closest to ours in diameter size and to the approach to use only the nodules marked by the radiologist, which identified the highest number of lesions in each exam, so from this work we achieved a superior result in comparison (AUC of 0.856 vs 0.820). Reeves et al. [Revees and et al 2006] used nodules a little bigger than ours and from different datasets, while Dhara et al. [Dhara and et al 2016] did not mention the diameter size of the nodules used from LIDC.

Table 4. Comparison with other studies in the classification of small lung nodules as benign or malignant.

Reference Authors	AUC	Classifier
Presented Work	0.856	SVM with Relief and EGA
Felix et. al. [Felix et al. 2016]	0.820	MLP with EGA
Reeves et. al. [Revees and et al 2006]	0.772	SVM with Radial Basis Funtion
Dhara et. al. [Dhara and et al 2016]	0.950	SVM

4. Conclusion

In this work we used 4 groups of features: 3D Texture Features, 3D Margin Sharpness Features, 3D Shape Features and 3D Intensity Features, leading to a total of 71 attributes. The classifiers we analysed were Support Vector Machine (SVM) and Multilayer Perceptron (MLP).

Working with SVM classifier together with Relief and Evolutionary Genetic Algorithm (EAG) showed to be the best choice to classify early pulmonary nodules. The use of EAG also showed the importance to improve the AUC and still reduce the dimensionality of the features helping in the cost and time of processing. The groups of attributes separately showed interesting results, but combining them improves the results when feature selection is performed with EAG and Relief. We also showed that, in a fair comparison between SVM and MLP using all features and EGA, SVM performs better.

References

Aerts, H. J. W. L. and et al (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. In *Nature communications*, volume 5, n.

4006, pages 1–8.

Akgul, C. B. and et. al. (2011). Content-based image retrieval in radiology: current status and future directions. *J Digit Imaging*, 24(2):208–222.

Alilou, M. and et al (2014). A Comprehensive Framework for Automatic Detection of Pulmonary Nodules in lung CT Images. *Image Analysis & Stereology*, 33(1):13–27.

Armato, S. G. and et al (2011). The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931.

Bartholmai, B. J. and et al (2015). Pulmonary nodule characterization, including computer analysis and quantitative features. *Journal of Thoracic Imaging*, 30(2):139–156.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computer and Electrical Engineering*, 40:16–28.

Choi, W. and Choi, T. (2013). Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. *Entropy*, 15:507–523.

Choi, W. and T., C. (2014). Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Computer Methods and Programs in Biomedicine*, 113:37–54.

Chuquicusma, M. and et al (2017). How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. *CoRR*, abs/1710.09762.

Ciompi, F. and et al (2017). Towards automatic pulmonary nodule management in lung cancer screening with deep learning. In *Scientific reports*.

Dhara, A. K. and et al (2016). A combination of shape and texture features for classification of pulmonary nodules in lung ct images. *Journal of Digital Imaging*, pages 1–10.

Dilger, S. K. and et al (2015). Improved pulmonary nodule classification utilizing lung parenchyma texture features. *Proc. SPIE*, 9414:94142T–10.

Felix, A., Juior, J., Oliveira, M., and Machado, A. (2016). Using 3d texture and margin sharpness features on classification of small pulmonary nodules. *Patterns amd Images (SIBGRAPI)*.

Ferreira-Junior, J. R., Oliveira, M. C., and de Azevedo-Marques, P. M. (2016). Cloud-based nosql open database of pulmonary nodules for computer-aided lung cancer diagnosis and reproducible research. *Journal of Digital Imaging*, pages 1–14.

Gillies, R., Kinahan, P., and Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577. PMID: 26579733.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621.

INCA (2020). Inca, `http://www.inca.gov.br/`. Acesssado em 08-02-2020.

junior, J., Oliveira, M., and Azevedo-Marques, P. (2016). Cloud-based nosql open database of pulmonary nodules for computer-aided lung cancer diagnosis and reproducible research. *J Digit Imaging*, 29(6):716–729.

Knight, S. and et, a. (2017). Progress and prospects of early detection in lung cancer. *Royal Society Journals*, 7(9):170070.

Lee, P. and et al. Rapidminer studio. Available in `https://rapidminer.com/products/studio/`.

Madero Orozco, H. and et al (2015). Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *BioMedical Engineering OnLine*, 14(1).

Neal, R., Sun, F., Herman, J., and Callister, M. (2019). Lung cancer. *BMJ*, 365.

Reeves, A., Xie, Y., and Jirapatnakul, A. (2015). Automated pulmonary nodule ct image characterization in lung cancer screening. *International Journal of Computer Assisted Radiology and Surgery*, 11(1):73–88.

Revees, A. P. and et al (2006). On measuring the change in size of pulmonary nodules. *IEEE Transactions on Medical Imaging*, 25(4):435–450.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Shewaye, T. N. and Mekonnen, A. A. (2016). Benign-malignant lung nodule classification with geometric and appearance histogram features. *CoRR*, abs/1605.08350.

Siegel, R. L., Miller, K. D., and DMV, A. J. (2017). Cancer statistics, 2017. *Journal CA CANCER*, 67(1):7–30.

Tammemagi, M. and Lam, S. (2014). Screening for Lung Cancer Using Low Dose Computed Tomography. *BMJ*, 348.

Wu, H. and et. al. (2013). Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography. *J Digit Imaging*, 26(4):797–802.

Xu, J. and et al (2012). Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval. *Medical Physics*, 39(9):5405–5418.

Yan, R. and et, a. (2018). The use of low-dose ct intra- and extra-nodular image texture features to improve small lung nodule diagnosis in lung cancer screening. In *2017 IEEE Nuclear Science Symposium and Medical Imaging Conference, NSS/MIC 2017 - Conference Proceedings*, pages 1–4. Institute of Electrical and Electronics Engineers Inc.

Zappa, C. and Mousa, S. (2016). Non-small cell lung cancer: current treatment and future advances. *Translational Lung Cancer Research*, 5(3):288–300.