Using 3D Texture and Margin Sharpness Features on Classification of Small Pulmonary Nodules

Ailton Felix, Marcelo Oliveira, Aydano Machado
Institute of Computing (IC)
Federal University of Alagoas (UFAL)
Maceió, Alagoas, Brazil
Email: {aflf,oliveiramc,aydano.machado}@ic.ufal.br

José Raniery
Ribeirao Preto Medical School (FMRP)
University of Sao Paulo (USP)
Ribeirão Preto, São Paulo, Brazil
Email: jose.raniery@usp.br

***Abstract*—The lung cancer is the reason of a lot of deaths on population around the world. An early diagnosis brings a most curable and simpler treatment options. Due to complexity diagnosis of small pulmonary nodules, Computer-Aided Diagnosis (CAD) tools provides an assistance to radiologist aiming the improvement in the diagnosis. Extracting relevant image features is of great importance for these tools. In this work we extracted 3D Texture Features (TF) and 3D Margin Sharpness Features (MSF) from the Lung Image Database Consortium (LIDC) in order to create a classification model to classify small pulmonary nodules with diameters between 3-10mm. We used three machine learning algorithm: k-Nearest Neighbor (k-NN), Multilayer Perceptron (MLP) and Random Forest (RF). These algorithms were trained by different set of features from the TF and MSF. The classification model with MLP algorithm using the selected features from the integration of TF and MSF achieved the best AUC of 0.820.**

***Keywords*-lung cancer; small nodules; early diagnosis; computer-aided diagnosis; texture features; margin sharpness features; classification; machine learning.**

I. Introduction

Cancer is characterized as an abnormal cell growth that invades and destroys neighboring tissues. Lung cancer is the most frequently diagnosed cancer and accounts for highest number of cancer-related deaths compared to any other cancer [1].

The survival rate for lung cancer analyzed in five years is only 15%. However, if the disease is identified at an early stage, the survival rate increases to 49%[2]. Therefore, early-stage nodules identification becomes a significant analysis in lung cancer screening; As soon as they are found more curable and simpler treatment options may be available [3]. Furthermore, nodule malignancy classification depends on temporal aspects like growth rate and size change between two time-separated Computed Tomography (CT) scans [4]. So, besides increasing the likelihood of patient survival, an early identification of potentially malignant pulmonary nodules also help emotionally the patient, avoiding the necessity to wait for days or months to measure the change in size, form or texture of the nodule.

Generally, nodules are visually evaluated and verbally characterized with a lexicon of radiologic features/descriptors and terms that are semi-quantitative but subjectively assessed such as: spiculated, smooth, flat, spherical, and others. This features may have a large amount of subjective, experiential and perceptual variability [5]. The medical image interpretation process has shown significant inter-observer variation in numerous studies due various aspects, e.g. time constraints, readers' perceptual errors, lack of training, or fatigue [6]. Moreover, small lung nodules (those less than 1cm in diameter) noted on CT images make the differential diagnosis clinically difficult and may confuse clinical decision-making [7], [8]. Because, among other factors, small pulmonary nodules have low contrast in comparison to the lung tissue and can be attached to other complex lung structures (Fig. 1)[9]. Therefore, the diagnosis of small lung nodules is a challenging task for specialists but very important to the patient's survival.

(a) 9.2mm vessel-connected. (b) 6.9mm vessel-connected (c) 5.7mm isolated

Fig. 1. Examples of pulmonary nodules (highlighted in red) with their size and anatomical structure connected to it.

In order to aid the radiologist in the medical images interpretation, Computer-Aided Diagnosis (CAD) tools have been used to determine the probability of malignancy of a lung nodule based on image features. Moreover, CAD tools have the potential to improve the accuracy of nodule classification (likely malignant or benign) by acting as a second opinion to specialists [10].

Several quantitative features have been used to characterize pulmonary nodules, e.g. texture, shape, form, density, etc [11], [12], [13]. In particular, some studies have used these quantitative features in the characterization of small pulmonary nodules [10], [3]. But the question whether these quantitative parameters are able to confer an advantage or not in the classification between benign and malignant nodules still remains [10]. So, there is a need to discovery relevant contents from the images to improve the performance of CAD systems and

there are few works using features on small lung nodules classification.

The objective of this work was to create a classification model for small pulmonary nodules, with diameter between 3-10mm, classifying them from benign and malignant using 3D Texture Features (TF) and 3D Margin Sharpness Features (MSF).

A. Related work

Reeves AP et al. [3] used 46 image features such as: 3D geometry features, 3D features of the density distribution, surface curvature features and features of the margin, to determine the malignancy status of pulmonary nodules evaluated with combined image data from the two large datasets, the Early Lung Cancer Action Program (ELCAP) and the National Lung Cancer Screening Trial (NLST), with a total of 736 nodules, 412 malignant and 324 benign with volumetric derived diameters between 3-29mm size-unbalanced. Different data subsets were used for such to determine the impact of class size distribution imbalance in datasets. One was the size-balanced nodule dataset, with 326 nodules (163 malignant and 163 benign) and volumetric derived diameters between 5-14mm. For classification were used: the distance weighted k-NN, the Support Vector Machine (SVM) with a Polynomial kernel (SVM-P), with a Radial Basis Function kernel (SMV-R), the Logistic Regression and the size threshold. With a 5-fold cross validation, a mean AUC of 0.772, with standard deviation of 0.031 was achieved with the SVM-R for the size-unbalanced data sets, the best performance. The best classification performance for the balanced dataset achieved average AUC of 0.708 (standard deviation 0.062) with SVM-P trained on balanced data.

Dhara et al. [14] used a set of 49 features combining 2D shape-based, 3D shape-based, 3D margin-based, 2D texture-based and 3D texture-based features on the classification of benign and malignant pulmonary nodules from 891 cases of Lung Image Database Consortium (LIDC) and Image Database Initiative public database. The classification scheme used different configurations of the databases regarding the classifications made by the radiologists. Using the SVM algorithm with a 5-fold cross validation approach their best AUC average performance achieved was 0.950. It is important to say that this work did not take into account nodule size issues on the classification.

B. Work structure

The remainder of this paper is organized as follows: section II describe implementation and details from the material and method used. Section III presents the results and discussion of this work. Lastly, section IV concludes this paper.

II. Materials and Methods

The overview schema of this work can be view on Fig. 2. Fist, we create a database (Section II-A) from a medical imaging repository (LIDC) in order to integrate information about exams data, images features and nodule size. Next, we had a step to measure the nodules size (Section II-B), we defined a size threshold to select the small nodules of our database (Section II-C), extracted the image features from small nodules selected (Section II-D), we carried out a feature selection step (Section II-E) and then we classified such nodules (Section II-F). The results of Nodule Size Measuring and Feature Extraction stage were stored in our database.

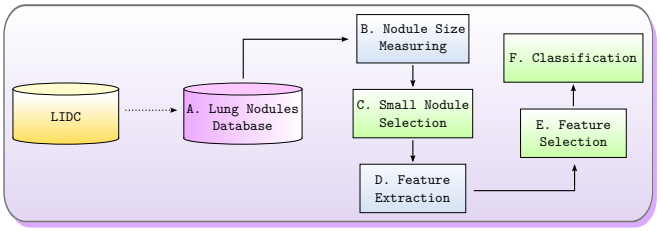


Fig. 2. General schema used in this work.

The feature selection and classification process were performed using the tool RapidMiner Studio [15], version 6.5.002. The tests were performed on a PC Intel Core i5, 3.10Hz CPU and 8GB RAM with operational system GNU/Linux Ubuntu 14.04 LTS.

A. Lung Nodules Database

We used the medical images from the LIDC [16], which consists of CT scans for lung cancer with lesions identified and classified by four experienced radiologists in a process of image interpretation which required the experts to read the CT scans and marking of lesions using a graphical interface. The identified nodules were ranked by radiologists according to subjective characteristics, among them likelihood of malignancy, following the conditions:

- Malignancy 1: high probability to be benign;
- Malignancy 2: moderate probability to be benign;
- Malignancy 3: indeterminate probability;
- Malignancy 4: moderate probability to be malignant;
- Malignancy 5: high probability to be malignant.

The LIDC is a collection not organized on database schema, so, there is no correlation between images, exams data and classification of nodules by radiologists. Furthermore, the LIDC does not contain information about nodule size or image features.

We created a database using a NoSQL approach Document-oriented [17], the Data Base Management System (DBMS) used was the MongoDB [18]. All lesions images were manually segmented using the radiologist's marks and then placed into our database. We extracted the image features from these lesions images. As the LIDC has four radiologist's marks, we use only one of the four marks to avoid redundancies. The criterion for choice was that made by the radiologist that identified the highest number of lesions in each exam.

Our Database has 752 exams and 1,944 lung nodules from LIDC on five ratings probability of malignancy. However, nodules with likelihood of malignancy 3 were discarded because they have probability of indeterminate malignancy, resulting

TABLE I
NODULE NUMBERS BETWEEN 3-10MM USED FROM OUR DATABASE.

	Benign		Malignant		
Likelihood of Malignancy	1	2	4	5	**Total**
Nodule Numbers	69	68	123	14	274
Sum	137		137		

in 1,171 nodules. For this work, nodules with probability of malignancy 1 and 2 were considered benign, and nodules with probability of malignancy 4 and 5 were considered malignant.

B. Nodule Size Measuring

The nodule size can be assessed as a single 2D measure of greatest diameter, typically performed in the axial plane along the axis of longest diameter [5]. Thus, for each nodule of the database we calculate the distance between the minimum and maximum coordinates in the respective x and y axes and choose the one with the longest distance.

C. Small Nodule Selection

The smallest nodule found in our database has 3.27mm in diameter. According to Bartholmai et al. [5], nodules <10mm have a nonzero risk for malignant and nodules greater than 10mm are much more likely to be malignant. Therefore, in order to prepare our classification model to face nodules as small as possible and to not work with nodules most likely to be malignant, we used the threshold diameter 3-10mm.

Due to the nodule diameter threshold used (3-10mm), our database provided a number of benign nodules much greater than malignant ones, which was expected because of the higher chances of small nodules to be benign [5], [7], [13]. However, in order to perform a fair classification, we balanced the number of benign and malignant cases, as presented in Table I.

D. Feature Extraction

The process of image feature extraction consists on removing of numeric values that represent the image visual content (images descriptors) through the implementation of algorithms [19]. After extraction of image descriptors, the features are stored in a feature vector. In this work, we used two categories of image features: 3D Texture Features and 3D Margin Sharpness Features.

Texture feature became particularly important due to its capacity to reflect details contained within a lesion in an image [6]. The variation of texture patterns of nodules provide strong indicators of its nature malignant or benign. For example, the presence of fat or calcification are strong indicators of a benign tumor and result in an irregular distribution of texture. On the other hand malignant nodules have uniform texture produced by the presence of necrosis [20], [21].

A margin sharpness feature is important to differentiate lesions in terms of potential malignancy because cancer tumors grow into neighboring tissues [22]. According to Xu et al. [23], a sharper margin will have a more abrupt transition and may have a higher difference of intensities outside and inside the lesion, whereas a blurred margin will have a smoother transition and may have a smaller intensity difference.

1) 3D Texture Features: We used Gray Level Co-occurrence Matrix (GLCM) to obtain texture attributes. GLCM is a technique to extract information from second-order statistical texture. It obtain, from a single image, the occurrence probability of a pixel pair with intensity i, j and spacing between the pixels of Δx and Δy in the dimensions x and y, respectively, given a distance d and orientation θ [24].

A 3D texture analysis applied to the calculation of GLCM in an image volume extends the probability of pairs of voxels to the Z-axis. Second-order statistics are applied to the GLCM producing the texture attributes. Haralick et al. [25] suggested the texture features used in this work, which are listed below:

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

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

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

$$\text{inertia} = \sum_{i,j} (i-j)^2 C(i,j), \tag{4}$$

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

$$\text{shade} = \sum_{i,j} (i+j-\mu_x-\mu_y)^3 C(i,j), \tag{6}$$

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

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

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

where $C(i,j)$ are the elements from the GLCM, μ_x and μ_y are the mean, σ_x and σ_y are the standard deviation, obtained by the following equations:

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

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

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

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

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

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

The TF vector was obtained by calculating the nine attributes (Equations 1-9) applied to the co-occurrence matrices performed in orientations $0°$, $45°$, $90°$ and $135°$, and distance of 1 voxel. In this case, each nodule was associated with a 36-dimension vector of TF.

2) 3D Margin Sharpness: A 3D margin sharpness analysis was also implemented in this work to characterize pulmonary nodules. The implementation was partially proposed by Xu et al. [23], in which the authors draw perpendicular lines over the borders on all nodule slices. The implementation is as follows: twenty control points were automatically selected on the marked lesion edge, starting by the first point marked by the specialist (Fig. 3(a)). If the boundary has p pixels, than a control point is marked every $\frac{p}{20}$ pixels. Normal lines were drawn at each of the 20 control points across the nodule boundary (Fig. 3(b)). A mask was created to eliminate the line segments that cross the lung wall because, otherwise, it will introduce pixel information that does not belong to the nodule or lung tissues. The mask was generated by applying a threshold algorithm along with morphological dilation operation in the original CT image (Fig. 3(c)). After excluding normal line segments that do not belong to the lung by means of the lung mask application (Fig. 3(d)), pixel intensities from the remaining line segments from all nodule images were recorded in a single sorted array. Then a data statistical analysis was performed by extracting statistical attributes from the pixel intensities sorted array. The MSF vector was composed by the statistical features listed in Equations 16-27, in which x is the pixel intensities array of size n, x_1 is the intensity value of a pixel outside the nodule and x_n is the intensity value of a pixel inside the nodule.

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

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

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

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

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

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

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

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

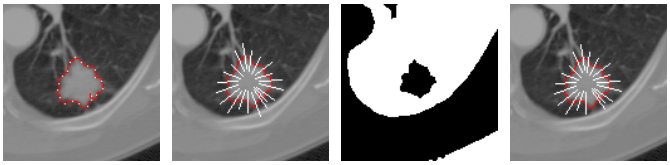
$$\text{standard deviation (s)} = \sqrt{\upsilon}, \tag{24}$$

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

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

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

Therefore, each nodule is characterized as a 12-dimension vector of MSF.



(a) Boundary control points. (b) Normal line segments. (c) Cropped mask obtained. (d) Cropped final output image.

Fig. 3. Output images from the 3D margin sharpness analysis.

3) Integration: Traina et al. asserts in [26] that texture and shape features should be integrated to provide better discrimination in the comparison process. Therefore, the texture and margin sharpness attributes were concatenated in order improve our classification model.

E. Feature Selection

A large number of features in a machine learning algorithm can leading a higher likelihood of noise or irrelevant features, hindering the learning process. This problem is known how curse of dimensionality [27]. To avoid this problem and to select the most relevant features on classification of small pulmonary nodules we applied a selection feature technique called Evolutionary Genetic Algorithm (EGA) [28] on TF, MSF and on integration of this categories of features.

The EGA is based on genetic and evolutionary theory, where the most environmentally adapted organisms are more likely to have their features reproduced in a new generation. Some of the advantages of genetic algorithms is the fact that they perform simultaneous searches in various regions of the solution space. This allows them to find various solutions, and makes it a global search method.

In the context of our work, our population is made up of individuals formed by binary vectors representing the presence/absence of a given feature. The selected individuals for reproduction were chosen using tournament criteria. In the reproductive phase, the chosen operators were: crossover and mutation, with applying probabilities to each individual 50% and 5%, respectively. The crossover type applied was one-point.

F. Classification

In order to build the classification model we used the k-Nearest Neighbor (k-NN) [29], an Artificial Neural Network

(ANN) technique called Multilayer Perceptron (MLP) [30] and Random Forest (RF) [31] machine learning algorithms. These techniques have been applied in both detection and classification of pulmonary nodules [32], [33], [34].

The classification model was evaluated with a 10-fold cross validation with the 274 small nodules selected. Three sets of features were separately used on each classifier: 3D Texture Features (TF), 3D Margin Sharpness Features (MSF) and Integration (I). For each set of features we evaluated also the classification performance with the selected features.

For the k-NN, k varied in the odd natural interval [1,15]. Two euclidean distance and correlation similarity metrics were used separately with each k value. With the MLP, we used 500 training cycles with 0.3 learning rate and 0.2 momentum, the performance with one and two hidden layers were evaluated. Ultimately, with the RF, we did tests with the generation of 50, 100, 150 and 200 trees, information gain was chosen as selection criteria with maximal depth 30. Pruning and prepruning were not applied. For each classifier, the best results achieved using this methodology were considered for comparison of the results.

III. Results and Discussion

We used the Area Under the ROC Curve (AUC) [35] to assess the performance of the classifiers on each set of features. The Tables II, III and IV present the classification results (mean $\pm$ standard deviation) over a 10-fold cross validation of each machine learning algorithm without and with feature selection (All Features and Selected Features on the Tables, respectively).

The classification model using TF achieved highest average AUC of 0.779 (σ = 0.087) with the k-NN algorithm using the selected features (Table II). All the machine learning algorithms had its performance improved using the selected features from TF. In particular, the k-NN used only 17 features from 36 TF.

The classification model using MSF obtained highest average AUC of 0.783 (σ = 0.077) with the k-NN algorithm using the selected features (Table III). All the machine learning algorithms had its performance improved using the selected features from MSF. In particular, the k-NN used 7 features from 12 MSF. So, the classification performance with TF and MSF was quite similar (difference between areas of 0.004).

The best results were achieved using TF and MSF integration and feature selection. The MLP algorithm obtained the highest average AUC of 0.820 (σ = 0.053) (Table IV). The MLP used 26 features (21 TF and 5 MSF) from 48 features. The results with k-NN and RF algorithms using integration with selected features outperformed the results showed on Tables II and III and both of them used TF and MSF on classification. This show that TF and MSF were both important for our classification model for small pulmonary nodules. The ROC curve showed on Fig. 4 confirms the superiority of MLP algorithm compared to the others best results using TF and MSF separately.

Comparing our work with Reeves AP et al. [3], we had a positive difference between the areas under the ROC curves of 0,048. [3] also took into account the diameter of the nodules to train and classify the machine learning algorithms. However, it is import to say that the image datasets used were different and the best result was achieved by a different machine learning algorithm, the SVM. Nevertheless, comparing our result with Dhara et al. [14], we obtained a negative difference between the areas under the ROC curves of 0,130, this can be explained because the authors used a more diverse set of 2D and 3D shape, margin and texture features. However, [14] did not take into account the diameter of the nodules, which eliminates some challenges that we faced by working with small nodules.

TABLE II
Small Nodule Classification using 3D Texture Features

	AUC	
	All Features	**Selected Features**
k-NN	0.675 ± 0.076	0.779 ± 0.087
MLP	0.736 ± 0.141	0.747 ± 0.086
RF	0.732 ± 0.094	0.758 ± 0.072

TABLE III
Small Nodule Classification using 3D Margin Sharpness Features

	AUC	
	All Features	**Selected Features**
k-NN	0.719 ± 0.091	0.783 ± 0.077
MLP	0.718 ± 0.057	0.758 ± 0.071
RF	0.705 ± 0.060	0.749 ± 0.100

TABLE IV
Small Nodule Classification using Integration

	AUC	
	All Features	**Selected Features**
k-NN	0.712 ± 0.063	0.804 ± 0.065
MLP	0.722 ± 0.087	0.820 ± 0.053
RF	0.771 ± 0.085	0.797 ± 0.086

A. Challenges

The small diameter of nodules that we are using (3-10mm), bring us a major challenge in classification stage due to the small amount of information (pixels) of the nodules. According to Reeves et al. [3], nodules of small size have less image information in CT images than large nodules due to the number of fixed-size image pixel elements (pixels) that they span. For example, a 2mm nodule spans in the order of 8 pixels, a 3mm nodule 27 pixels, a 4mm nodule 64 pixels and a 5mm nodule 620 pixels; further, for all these cases, a large majority of these pixels are partial pixels; that is, they consist of a mixture of the nodule tissue and the surrounding lung tissue. It is important to say that texture and margin sharpness features use internal information of the nodule.

Using the diameter threshold between 3-10mm our database has only 14 nodules with malignancy 5 against 123 with

malignancy 4. It was just this total number used in our study, as can be seen in Table I. Remember that nodules with malignancy 5 indicate high probability to be malignant, this way, it is possible to assume that these nodules have more characteristics of a malignant nodule that nodules with malignancy 4. Therefore, as the learning process on malignant nodules by the classifier is practically performed on the nodules with malignancy 4, due to the discrepancy amount of nodules compared with malignancy 5, this process is impaired, consequently hindering the classification process.

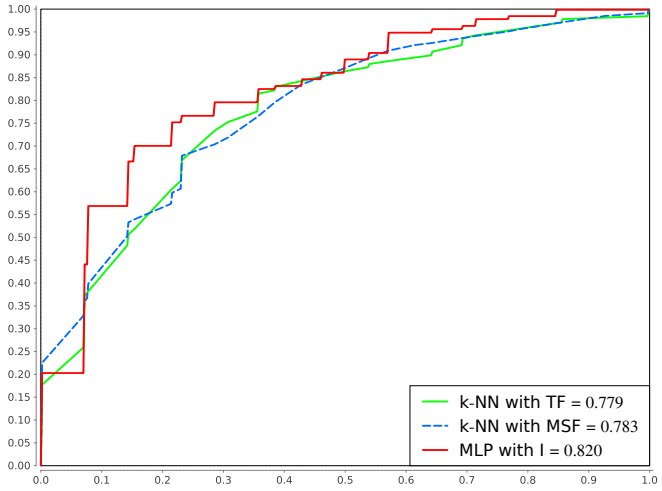


Fig. 4. Comparison of ROC curve among the best results of classification models using TF, MSF and I.

IV. Conclusion

In function of the results obtained, with the machine learning algorithm and features used, the best classification model for small pulmonary nodules must use the MLP algorithm with texture and margin sharpness features integrated and adopt the following set of features: sum of logs, arithmetic mean, geometric mean, population variance, standard deviation, energy 0°, inertia (0°, 45° and 135°), homogeneity (0°, 45° and 135°), correlation (0° and 45°), shade (0°, 45° and 135°), promenance (0°, 90° and 135°), variance (0° and 90°), idm (0°, 45° and 90°) and entropy 45°.

The classification model using TF and MSF separately have a similar performance in the classifications of small pulmonary nodules using k-NN, MLP and RF machine learning algorithms. The k-NN algorithm achieved the best performance in both scenarios.

Our classification model for small pulmonary nodules still has underperforming compared to state of the art. In order to improve our model, as future work we plan to use more machine learning algorithms and to include in our set of features the lung parenchyma surrounding the nodule, that in the work Dilger et al. [10] it proved quite promising to include information that increases amount of data available, which attacks just our challenge of the number of pixels that a small nodule has. Advances in this area are important since the early nodules classification is challenging the expert, but critical to patient survival.

Acknowledgment

The authors would like to thank the financial supporting provided by the brazilian institution Coordination for the Improvement of Higher Education Personnel (CAPES).

References

[1] A. Jemal, P. Vineis, F. Bray, L. Torre, and D. Forman, *The Cancer Atlas*, 2nd ed. American Cancer Society, 2014. [Online]. Available: www.cancer.org/canceratlas

[2] P. Aggarwal, H. Sardana, and R. Vig, "Content based image retrieval approach in creating an effective feature index for lung nodule detection with the inclusion of expert knowledge and proven pathology," C. M. I. Reviews, Ed., vol. 10. Current Medical Imaging Reviews, 2014, pp. 178–204.

[3] A. P. Reeves, Y. Xie, and A. Jirapatnakul, "Automated pulmonary nodule ct image characterization in lung cancer screening," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 1, pp. 73–88, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11548-015-1245-7

[4] A. P. Reeves, A. B. Chan, D. F. Yankelevitz, C. I. Henschke, B. Kressler, and W. J. Kostis, "On measuring the change in size of pulmonary nodules," *IEEE Transactions on Medical Imaging*, vol. 25, no. 4, pp. 435–450, April 2006.

[5] B. J. Bartholmai, C. W. Koo, G. B. Johnson, D. B. White, S. M. Raghunath, S. Rajagopalan, M. R. Moynagh, R. M. Lindell, and T. E. Hartman, "Pulmonary nodule characterization, including computer analysis and quantitative features," *Journal of Thoracic Imaging*, vol. 30, no. 2, pp. 139–156, March 2015.

[6] C. B. Akgül, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar, "Content-based image retrieval in radiology: Current status and future directions," *Journal of Digital Imaging*, vol. 24, no. 2, pp. 208–222, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10278-010-9290-9

[7] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OncoTargets and Therapy 2015:8 20152022*. [Online]. Available: http://doi.org/10.2147/OTT.S80733

[8] D. F. Yankelevitz, R. Gupta, B. Zhao, and C. I. Henschke, "Small pulmonary nodules: Evaluation with repeat ctpreliminary experience," *Radiology*, vol. 212, no. 2, pp. 561–566, 1999, pMID: 10429718. [Online]. Available: http://dx.doi.org/10.1148/radiology.212.2.r99au33561

[9] M. Alilou, V. Kovalev, E. Snezhko, and V. Taimouri, "A comprehensive framework for automatic detection of pulmonary nodules in lung ct images," *Image Analysis & Stereology*, vol. 33, no. 1, pp. 13–27, 2014. [Online]. Available: http://www.ias-iss.org/ojs/IAS/article/view/1081

[10] S. K. Dilger, A. Judisch, J. Uthoff, E. Hammond, J. D. Newell, and J. C. Sieren, "Improved pulmonary nodule classification utilizing lung parenchyma texture features," *Proc. SPIE*, vol. 9414, pp. 94 142T–94 142T–10, 2015. [Online]. Available: http://dx.doi.org/10.1117/12.2081397

[11] W. J. Choi and T. S. Choi, "Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, pp. 37–54, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2013.08.015

[12] M. C. Oliveira and J. R. Ferreira, "A bag-of-tasks approach to speed up the lung nodules retrieval in the bigdata age," in *e-Health Networking, Applications Services (Healthcom), 2013 IEEE 15th International Conference on*, Oct 2013, pp. 632–636.

[13] Y.-X. J. Wang, J.-S. Gong, K. Suzuki, and S. K. Morcos, "Evidence based imaging strategies for solitary pulmonary nodule," *Journal of Thoracic Disease*, vol. 6, no. 7, p. 872, 2014.

[14] A. K. Dhara, S. Mukhopadhyay, A. Dutta, M. Garg, and N. Khandelwal, "A combination of shape and texture features for classification of pulmonary nodules in lung ct images," *Journal of Digital Imaging*, pp. 1–10, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10278-015-9857-6

[15] P. Lee, I. Mierswa, L. Bauerle, B. Doyle, T. McHugh, F. Gedling, T. Wentworth, and M. Mierswa. Rapidminer studio. Last accessed 04-04-2016. [Online]. Available: https://rapidminer.com/products/studio/

[16] S. G. Armato, G. McLennan, L. Bidaut *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011. [Online]. Available: http://scitation.aip.org/content/aapm/journal/medphys/38/2/10.1118/1.3528204

[17] C. Strauch, *NoSQL databases*. Stuttgart Media University, 2011.

[18] S. Tiwari, *Professional NoSQL*. John Wiley and Sons, Inc, 2011.

[19] M. P. da Silva, "Processing similarity queries in medical images to the perceptual recovery guided by the user," Ph.D. dissertation, University of São Paulo (USP), 2009.

[20] J. J. Erasmus, J. E. Connolly, H. P. McAdams, and V. L. Roggli, "Solitary pulmonary nodules: Part i. morphologic evaluation for differentiation of benign and malignant lesions," *RadioGraphics*, vol. 20, no. 1, pp. 43–58, 2000, pMID: 10682770. [Online]. Available: http://dx.doi.org/10.1148/radiographics.20.1.g00ja0343

[21] S. Takashima, S. Sone, F. Li, Y. Maruyama, M. Hasegawa, and M. Kadoya, "Indeterminate solitary pulmonary nodules revealed at population-based ct screening of the lung: using first follow-up diagnostic ct to differentiate benign and malignant lesions," *American Journal of Roentgenology*, vol. 180, no. 5, pp. 1255–1263, 2003.

[22] J. E. Levman and A. L. Martel, "A margin sharpness measurement for the diagnosis of breast cancer from magnetic resonance imaging examinations," *Academic Radiology*, vol. 18, no. 12, pp. 1577 – 1581, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1076633211003904

[23] J. Xu, S. Napel, H. Greenspan, C. F. Beaulieu, N. Agrawal, and D. Rubin, "Quantifying the margin sharpness of lesions on radiological images for content-based image retrieval," *Medical Physics*, vol. 39, no. 9, pp. 5405–5418, 2012. [Online]. Available: http://scitation.aip.org/content/aapm/journal/medphys/39/9/10.1118/1.4739507

[24] M. C. Oliveira, W. Cirne, and P. M. de Azevedo Marques, "Towards applying content-based image retrieval in the clinical routine," *Future Generation Computer Systems*, vol. 23, no. 3, pp. 466–474, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X06001348

[25] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov 1973.

[26] A. J. M. Traina, A. G. R. Balan, L. M. Bortolotti, and C. Traina, "Content-based image retrieval using approximate shape of objects," in *Computer-Based Medical Systems, 2004. CBMS 2004. Proceedings. 17th IEEE Symposium on*, June 2004, pp. 91–96.

[27] J. E. Mason, M. Shepherd, J. Duffy, V. Keselj, and C. Watters, "An n-gram based approach to multi-labeled web page genre classification," *2014 47th Hawaii International Conference on System Sciences*, vol. 0, pp. 1–10, 2010.

[28] A. Rozsypal and M. Kubat, "Selecting representative examples and attributes by a genetic algorithm," *Intell. Data Anal.*, vol. 7, no. 4, pp. 291–304, Sep. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=1293868.1293870

[29] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, Nov. 1998. [Online]. Available: http://doi.acm.org/10.1145/293347.293348

[30] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 683–697, Sep 1992.

[31] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[32] S. T. Namin, H. A. Moghaddam, R. Jafari, M. Esmaeil-Zadeh, and M. Gity, "Automated detection and classification of pulmonary nodules in 3d thoracic ct images," in *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*, Oct 2010, pp. 3774–3779.

[33] J. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for {CT} images," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, pp. 202 – 209, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260713003532

[34] A. Tartar, N. Kl, and A. Akan, "A new method for pulmonary nodule detection using decision trees," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, July 2013, pp. 7355–7359.

[35] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *ReCALL*, vol. 31, no. HPL-2003-4, pp. 1–38, 2004.