

ANÁLISE FILOGENÉTICA PARA DIFERENCIAÇÃO ENTRE NÓDULOS MALIGNOS E BENIGNOS

Otília de S. Santos¹, Thayane de O. Simões¹, Laércio N. Mesquita¹, Alcilene D. de Sousa¹,
Antônio O. de C. Filho¹

¹Campus Senador Helvídio Nunes de Barros - Universidade Federal do Piauí (UFPI), Brasil

Resumo: Apresenta uma metodologia de auxílio no diagnóstico por computador (*Computer-Aided Diagnosis -CADx*) para classificação de malignidade ou benignidade dos nódulos pulmonares em imagens de tomografia computadorizada. O uso dos índices de diversidade filogenética para extração das características dos nódulos, a classificação é realizada com a ferramenta WEKA usando múltiplos classificadores, validação dos resultados com as métricas kappa, *Area Under the Curve*, sensibilidade, especificidade e acurácia. Os testes mostraram resultados bem objetivos e robustos para uma metodologia CADx com uma acurácia de 98,1%, sensibilidade 98,7%, especificidade 97,9%, um kappa de 0,95 e uma *Area Under the Curve* de 0,99. Os resultados obtidos comprovaram o bom desempenho das técnicas de extração de características de textura através dos índices apresentados, com uma precisão de 98,1%.

Palavras-chave: Imagens médicas, Diagnóstico pulmonar, Índices de diversidade.

Abstract: Present a methodology to assist in the computer diagnosis (*Computer-Aided Diagnosis -CADx*) to classify pulmonary nodules in malignant and benign in CT images. Using phylogenetic diversity index to extract the characteristics of the nodes, the classification made with WEKA tool, validating the results with the following metrics kappa, ROC curve, sensitivity, specificity and accuracy. The tests showed very accurate and robust results for integration in a CADx tool with an accuracy of 98.1%, 98.7% sensitivity, 97.9% specificity, a kappa of 0.95 and an AUC of 0.99. The results indicated a good performance of texture extraction techniques through the indexes presented with an accuracy of 98.1%.

Keywords: Tomography, Diagnosis, Diversity Index.

Introdução

O câncer é uma patologia com localizações e aspectos clínico-patológicos múltiplos e alguns tipos não possui sintomas, podendo ser detectado em vários estágios de evolução. Isso influencia na avaliação do seu diagnóstico, tornando-o mais dificultoso, como também na afirmativa de que a suspeita de câncer pode surgir diante dos sintomas mais variados possíveis, como a tosse e o sangramento pelas vias respiratórias ¹.

O câncer de pulmão está associado ao consumo de tabaco em 90% dos casos diagnosticados e tem uma taxa de crescimento anual de 2% em sua incidência no mundo todo. Comparados com os não fumantes, os tabagistas têm cerca de 20 a 30 vezes mais risco de desenvolver o câncer¹. Esse tipo de câncer apresenta a maior taxa de mortalidade e tem uma das menores taxas de sobrevivência após o diagnóstico.

Para auxiliar o especialista na busca, identificação de nódulos e alterações em imagens tomográficas, são desenvolvidos sistemas que utilizam o Processamento Digital de Imagens (PDI), que consiste

em um conjunto de técnicas para manipulação de imagens com a auxílio de um computador afim de facilitar a extração de informações presentes nessas imagens.

Os sistemas de Detecção e Diagnóstico auxiliados por computador (*Computer-Aided Detection (CAD) / Computer-Aided Diagnosis (CADx)*) ajudam os especialistas a lidar com um grande volume de informações dos pacientes e fornecer um diagnóstico preciso. Segundo Sousa², os sistemas CAD auxiliam na detecção de anormalidades, mas não realizam quaisquer tipos de diagnósticos sobre as mesmas. Os sistemas CADx, por sua vez, classificam as estruturas detectadas com anormalidade nas classes benignas ou malignas. Isto aumenta o grau de exatidão na detecção e diagnóstico, oferecendo uma segunda opinião ao especialista.

Esse trabalho conta com contribuições diretas em algumas áreas. Na área médica, a contribuição está no desenvolvimento de um sistema para CADx através da análise da textura do nódulo pulmonar. Na área da computação, a contribuição se dá nos seguintes aspectos: a) utilização de medidas de texturas baseadas nos índices de *phylogenetic diversity*³, *sum of phylogenetic distances*³, *phylogenetic species variability*³, *phylogenetic species richness*³ e *Mean nearest neighbor distance*⁴, e b) uso de árvores filogenéticas para caracterização dos nódulos pulmonares.

Trabalhos Relacionados

Na literatura, existem diversos trabalhos relacionados ao desenvolvimento de sistemas automáticos para diagnóstico do câncer de pulmão. Para esta finalidade, utilizam-se características extraídas de imagens médicas, objetivando a classificação em nódulos malignos ou benignos.

Froner⁵ apresentou uma avaliação para a utilização de dados de pacientes e atributos quantitativos de nódulos pulmonares em imagens de Tomografias Computadorizadas (TC) de pulmão para a construção de um modelo de classificação em termos de malignidade. Neste trabalho foi desenvolvido um modelo que melhor classifica a malignidade dos nódulos, seu melhor resultado tem uma curva ROC de 0,923 de precisão.

Orozco et al.⁶ apresentaram um trabalho que avaliou os descritores mais significativos utilizando o classificador SVM. O trabalho validou imagens de 45 TC das bases *Early Lung Cancer Action Program (ELCAP)* e *Lung Image Database Consortium – (LIDC-IRDI)*. Foram obtidos resultados significativos e promissores, com a acurácia de 82%; a sensibilidade de 90,9% e a especificidade de 73,91%.

Jhilam Mukherjee et al.⁷ desenvolveram um método que detecta e classifica nódulos pulmonares solitários a partir de imagens de TC. Esse método reduz a variabilidade nas detecções por segmentação automática e classificação de nódulos, seus resultados experimentais foram promissores no que diz respeito à classificação como nódulos malignos ou benignos.

Dandil et al.⁸ desenvolveram uma técnica na qual utilizam imagens TC para diferenciação entre tumores malignos e benignos. O sistema CADx projetado forneceu segmentação de nódulos usando *Self-Organizing Maps* e realiza a classificação entre nódulos benignos e malignos. Por fim, os resultados obtidos têm 90,63% de acurácia, sensibilidade de 92,30% e 89,47% de especificidade.

Han et al.⁹ apresentaram um trabalho que avaliou o desempenho de duas dimensões (2D) e características de textura 3D a partir de imagens de TC em nódulos pulmonares usando o banco de dados grande LIDC-IDRI num total de 905 nódulos, sendo 422 malignos e 483 benignos, com os resultados característicos na base de características de textura 3D de Haralick.

Os trabalhos relacionados expostos acima mostram que as metodologias baseadas em características de textura descrevem bem padrões em imagens, cujas medidas estatísticas são bastante utilizadas. Os resultados são promissores para o auxílio na detecção de câncer de pulmão, pelo valor da acurácia. Neste trabalho, pretende-se apresentar melhorias na descrição de padrões de textura das imagens de

TC, com aplicação do índice de diversidade filogenética e múltiplos classificadores para discriminar bem os nódulos pulmonares malignos e benignos.

Métodos

Para identificar os nódulos pulmonares em maligno e benigno, foi empregada a seguinte metodologia: utilização da base de imagens pública LIDC-IDRI; para a extração de características foi utilizado o descritor baseado na textura utilizando os índices de diversidade filogenética *Phylogenetic Diversity* (PD), *Sum of Phylogenetic Distances* (SPD), *Mean Nearest Neighbor Distance* (MNND), *Phylogenetic Species Variability* (PSV) e *Phylogenetic Species Richness* (PSR); para a classificação utilizou-se *Random Forest*¹⁰, *Random Tree*, *MultiLayerPerceptron*¹¹ e *JRip* e, por fim, a validação dos resultados. A Figura 1 apresenta um resumo das etapas seguidas pela metodologia proposta.

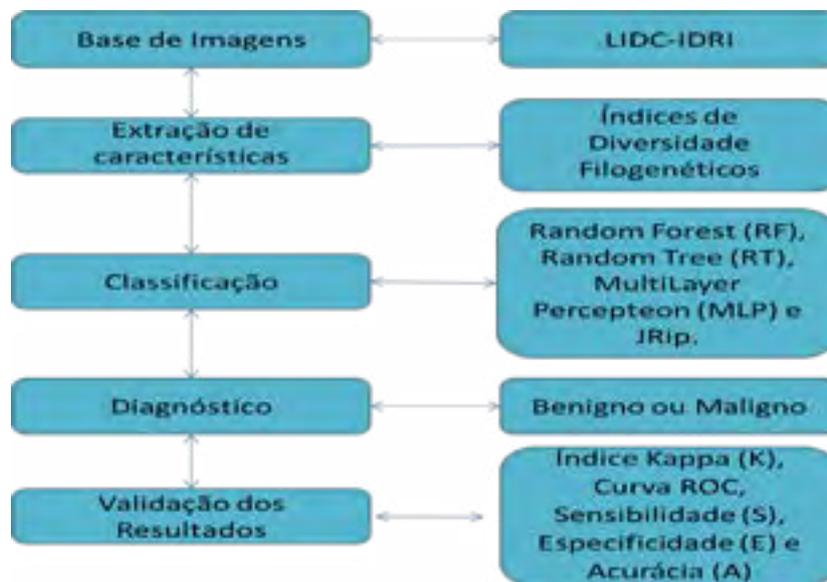


Figura 1: Método Proposto.

Base de Imagem - A Tomografia Computadorizada é um exame que permite a obtenção de imagens de cortes do corpo do paciente, sendo bastante utilizada como um exame médico de diagnóstico por imagem. Ela é excelente para detecção de alterações agudas ou crônicas no parênquima pulmonar.

O resultado visual da TC é monocromático, ou seja, são mostrados apenas os vários níveis de cinza, indo do totalmente preto ao branco, todas as estruturas presentes na TC puderam ser proporcionalmente quantificadas em unidades de densidade relativas ao padrão, chamadas de Unidades Hounsfield (UH).

Um nódulo é uma pequena massa de tecido que geralmente se forma em resposta às lesões, podendo ser benigno ou maligno. Um nódulo benigno é um tumor que surge e não se espalha para outras partes do corpo, tende a crescer mais lentamente e causa menos problemas à saúde do que um nódulo maligno¹², como demonstra a Figura 2.

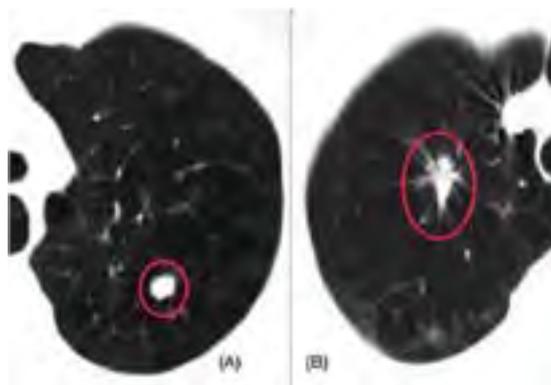


Figura 2: Exemplos de TC, onde é possível verificar o aparecimento de nódulos pulmonares: (A) TC com o nódulo benigno e (B) TC com nódulo maligno¹³.

No desenvolvimento desse trabalho utilizou-se a base de imagens LIDC-IDRI¹³, por ser uma base que preserva a identidade dos pacientes, como também apresenta a avaliação dos exames realizada por quatro especialistas. A base é um recurso internacional acessível via web para o desenvolvimento, treinamento e diagnóstico de sistemas especialistas que realizam o diagnóstico do câncer de pulmão.

A base LIDC-IDRI é disponibilizada pelo *National Cancer Institute of EUA* (NCI), sendo resultado da associação entre o consórcio de instituições conhecido LIDC-DRI. A base contém 1018 exames com uma quantidade variável de fatias por exame e cada exame contém um arquivo XML (*eXtensible Markup Language*) contendo as marcações e as avaliações de 4 especialistas¹⁴.

Índices de Diversidade Filogenética - A definição de textura encontrada na literatura é descrita de diversas formas. Segundo Haralick et al.¹⁵, a textura é definida como a característica de uma região relacionada a coeficientes de uniformidade, densidade, aspereza, regularidade, intensidade, dentre outras características da imagem. A textura será utilizada como base para a extração das características dos nódulos pulmonares.

Para Magurran¹⁶ a diversidade filogenética é a medida de uma comunidade que incorpora as relações filogenéticas das espécies. A forma mais simples da aplicação do índice de diversidade em imagens consiste na imagem que representa a comunidade ou região da mesma¹⁷, como apresentado na Figura 3.

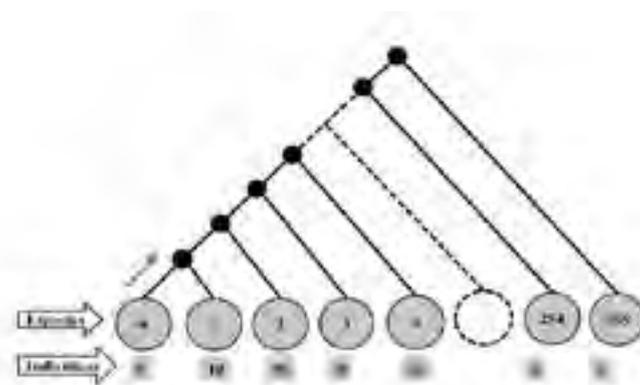


Figura 3: Árvore enraizada na forma de cladograma inclinado

Fonte: Oliveira¹⁸

Utilizamos os índices PD, SPD, MNND, PSV e PSR com a finalidade de localizar padrões em regiões de imagens de Tomografias. Nesse sentido os índices foram utilizados para descrever a textura

dos nódulos pulmonares extraídos dos exames de TC. Para a aplicação dos índices foram usadas as informações o nódulo pulmonar como: UH, quantidade de *voxel* de cada UH e similaridade entre duas UH.

Os índices de diversidade filogenética extraídos das árvores filogenéticas são empregados na biologia para comparar amostras de comportamento entre as espécies de diferentes áreas, assim como na área da computação para diagnosticar qual o padrão de benignidade ou malignidade de cada nódulo. Por isso, faz-se uma correlação entre a biologia e a metodologia proposta, conforme mostrado na Tabela 1.

Tabela 1. Correspondência entre os termos da biologia e o trabalho.

Biologia	Metodologia
Comunidade	Regiões de interesse (Nódulo) da imagem TC
Espécies	Número de HU da região
Indivíduos	Quantidade de voxel de cada UH
Distância Filogenética	Número de arestas entre duas espécies

O índice de diversidade filogenética *Phylogenetic Diversity* (PD) é obtido na biologia pela soma dos comprimentos dos braços da árvore filogenética das espécies de uma comunidade, segundo o autor Faith¹⁹. Quanto mais ramos a árvore filogenética possuir, mais distintos serão os grupos taxonômicos. Portanto, o PD de uma comunidade é a soma de todos os comprimentos dos ramos da porção de uma árvore filogenética conectando o conjunto focal das espécies, conforme apresenta a Equação 1,

$$PD = \frac{B \cdot \sum_{i=1}^S L_i A_i}{\sum_{i=1}^S A_i} \quad (1)$$

onde B é o número de ramificações da árvore, S é o número de espécies do conjunto focal, L_i é o comprimento do ramo i e A_i representa a abundância média de espécies (indivíduos) que compartilham i.

Para Webb, o índice *Mean Nearest Neighbor Distance* (MNND) é a distância filogenética média do parente mais próximo de todas as espécies, além disso, é equivalente às taxas de espécies por gênero. Esse índice pode ser calculado a partir da média ponderada da distância filogenética de cada vizinho mais próximo das espécies, com pesos iguais a abundância das espécies, conforme a Equação 2,

$$MNND = \sum_{m,n}^S \min(d_{mn}) a_m \quad (2)$$

onde S é número de espécies do conjunto focal, m e n são espécies, d_{mn} é a distância filogenética entre m e n e a_m é a abundância das espécies de m.

Segundo Helmus et al., o índice de diversidade filogenética *Sum of Phylogenetic Distances* (SPD) é a soma das distâncias filogenéticas entre cada par de espécies multiplicada pela distância filogenética média de cada par de espécies no conjunto focal, como apresentada na Equação 3,

$$SPD = \frac{\left(\frac{S(S-1)}{2}\right) \cdot \sum_{m < n}^S d_{mn} a_m a_n}{\sum_{m < n}^S a_m a_n} \quad (3)$$

onde S é número de espécies do conjunto focal, m e n são espécies, d_{mn} é a distância filogenética entre m e n, a_m é a abundância das espécies de m e a_n é a abundância das espécies de n.

No trabalho desenvolvido por Helmus et al., o índice *Phylogenetic Species Variability* (PSV) tem a variabilidade quantifica o parentesco filogenético, diminuindo a variação das características compartilhadas por todas as espécies da comunidade. A Equação 4 define a variabilidade de espécies filogenéticas, que resume o grau em que as espécies em uma comunidade são filogeneticamente relacionadas,

$$PSV = \frac{nrc - \sum c}{n(n-1)} = 1 - \epsilon \quad (4)$$

onde representa-se a soma dos valores da diagonal de uma matriz C, o somatório de todos os valores da matriz, n é o número de espécies e ϵ é a média dos elementos da diagonal de C.

O *Phylogenetic Species Richness* (PSR) quantifica o número de espécies em uma comunidade como nota-se nos parâmetros na Equação 5, onde o valor do PSR é a multiplicação do número espécies n pela variabilidade da comunidade.

$$PSR = n * PSV \quad (5)$$

Classificação - Segundo Mitchell et al.²⁰, a área de Aprendizagem de Máquina (AM) lida com o estudo de métodos computacionais que permite programas de computadores obterem melhorias na execução de tarefas de forma autônoma, por meio de experiências.

A classificação foi realizada pelo software Waikato Environment for Knowledge Analysis (WEKA), que é um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Contém ferramentas para pré-processamento de dados, classificação, regressão, clustering, regras de associação e visualização. Ele foi desenvolvido por um grupo de pesquisadores da Universidade de Waikato, Nova Zelândia ²¹.

A classificação é um método de reconhecimento automático de objetos, sendo que as entradas são as características extraídas das imagens TC do pulmão. Os classificadores testados usaram o método k-fold cross validation para obter os resultados. Os dados foram divididos em 10 conjuntos, sendo 9 deles para treinamento e 1 para testes. Este processo é repetido 10 vezes, de forma que o conjunto escolhido para o teste será diferente do anterior e no final é gerada uma média dos resultados.

O diagnóstico é feito a partir dos conjuntos de características previamente extraídas das imagens, as quais são submetidas à avaliação de um classificador que, a partir de um treinamento prévio, informa se o nódulo possui natureza maligna ou benigna.

Validação dos Resultados - Na validação dos resultados, utilizou-se as métricas de avaliação baseadas em estatísticas, como a Área sob a Curva ROC, que mensura o quanto o algoritmo é eficiente, usualmente referida em *Area Under the Curve* (AUC)²², Sensibilidade (S)²³, Especificidade (E)²³, Acurácia (A)²⁴ e índice Kappa (K)²².

Quando se avalia um teste de classificação, deve-se levar em consideração quatro perspectivas possíveis de ocorrências: Verdadeiro Positivo (VP), quando os nódulos são classificados corretamente como doentes (maligno); Falso Positivo (FP), que é o número de imagens que são classificadas erradamente como não doentes; Verdadeiro Negativo (VN), o número de imagens que são classificadas corretamente como saudáveis (benigno) e Falso Negativo (FN), o número de imagens que são classificadas erradamente como não saudáveis.

Segundo Martinez et al.²³ a sensibilidade é definida como a probabilidade do teste em fornecer um resultado positivo, desde que o indivíduo realmente seja portador da enfermidade. A especificidade, por sua vez, é definida como a probabilidade do teste em fornecer um resultado negativo, em que o indivíduo está livre da enfermidade. Já a Acurácia, para Metz²³, é a métrica que calcula o total de acertos em relação a todas as instâncias classificadas corretamente.

O índice Kappa (K), de acordo com Landis e Koch²², é uma medida de concordância que pode ser formulada para medir o desacordo de um conjunto de respostas, baseada em pesos, a qual mede a concordância entre um número de respostas baseando-se em observadores, chegando, assim, a um consenso.

Resultados e Discussão - Para os testes realizados neste trabalho, utilizou-se a base LIDC-IDRI utilizando 1402 nódulos, sendo 1009 benignos e 393 malignos. As características foram extraídas a partir dos índices de diversidade filogenética. Os classificadores *Random Forest* (RF), *Random Tree* (RT), *MultiLayerPerceptron* (MLP) e *JRip*, usados para classificar em maligno e benigno os nódulos pulmonares, os parâmetros dos classificadores com os valores padrões, através do software WEKA na versão 3.9. A classificação foi realizada com validação cruzada de k-folds, sendo k = 10.

Tabela 2. Resultados para classificação em benigno e maligno utilizando Índices de diversidade filogenética

	VP	FP	VN	FN	A	S	E	AUC	K
MLP	345	48	977	32	94,3%	91,5%	95,3%	0,959	0,856
RF	371	22	1004	5	98,1%	98,7%	97,9%	0,994	0,951
RT	370	23	993	16	97,2%	95,9%	97,7%	0,963	0,930
JRip	366	27	998	11	97,3%	97,1%	97,4%	0,962	0,932

A Tabela 2 indica que tanto o indicador do *kappa* quanto a acurácia obtiveram resultados promissores. O classificador *RF* obteve o melhor resultado com acurácia de 98,1%, sensibilidade de 98,7% e especificidade de 97,9%, tendo um *kappa* de 0,951 e uma AUC chegando ao resultado de 0,994.

Tabela 3. Comparação entre trabalhos relacionados e a metodologia proposta

Trabalhos	Base	Sensibilidade	Especificidade	Acurácia
Orozco <i>et al.</i> ⁽¹⁷⁾	ELCAP and LIDC	90,9%	73,91 %	82%
Dandil <i>et al.</i> ⁽⁴⁾	Privado	92,30 %	89,47%	90,63%
Metodologia	LIDC-IDRI	98,7%	97,9%	98,1%

De acordo com a Tabela 3, em comparação com os trabalhos relacionados, a metodologia proposta apresenta equilíbrio entre as três métricas (sensibilidade, especificidade e acurácia). Assim, em termos qualitativos, a metodologia apresenta um ótimo resultado para detecção de nódulos pulmonares em malignos ou benignos.

Conclusão e Trabalhos Futuros

O diagnóstico de câncer de pulmão auxiliado por computador é um tema relevante atualmente, que contribui de forma significativa para um diagnóstico correto e em menor tempo. O diagnóstico precoce representa um considerável aumento na probabilidade de sobrevivência dos pacientes. O presente trabalho apresentou uma metodologia automática para classificação de nódulos pulmonares baseada na análise de textura, extraída com os índices de diversidade filogenéticos e classificado através de

múltiplos classificadores, com intuito de identificar a natureza dos nódulos pulmonares em maligno ou benigno, sendo, assim, uma segunda opinião para o especialista da área.

Os resultados obtidos confirmaram um bom desempenho das técnicas de extração de textura através dos índices apresentados, com uma taxa de acerto de 98,1%.

Para trabalhos futuros, serão desenvolvidos e implementados outros índices de diversidade filogenética utilizando técnicas de aprendizados de máquinas, visando obter maiores taxas de acertos para a classificação das regiões de interesse. Pretende-se utilizar outras bases de imagens para verificar a eficácia da metodologia para classificação de malignidade ou benignidade dos nódulos pulmonares.

Agradecimentos

Os autores agradecem a Universidade Federal do Piauí (UFPI) e a Fundação de Amparo à Pesquisa do Estado do Piauí (FAPEPI).

Referências

- [1] Instituto Nacional de Câncer – INCA. O que é o câncer. 2016 Fev [citado 2016 fev 22]. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/cancer/site/oquee>.
- [2] Sousa JRFS. Metodologia para detecção automática de nódulos pulmonares [Dissertação]. São Luís: Universidade Federal do Maranhão; 2011.
- [3] Helmus MR, Bland TJ, Williams CK, Ives AR. Phylogenetic measures of biodiversity. *American Naturalist*. 2007;169, E68-E83.
- [4] Webb C O. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*. 2000;145-155.
- [5] Froner APP. Caracterização de nódulos pulmonares em imagens de tomografia computadorizada para fins de auxílio ao diagnóstico [Dissertação]. Porto Alegre: Pontifícia Universidade Católica do Rio Grande Do Sul Faculdade de Engenharia Elétrica; 2015.
- [6] Orozco HM.; Villegas OOV.; Sanchez, VGC.; Dominguez, HDJO.; Alfaro, MDJN. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *BioMedical Engineering OnLine* 20; 14: 9.
- [7] Mukherjee J., Choudhury SA.K., Kolkata, IA. Chakrabarti SH. Shaikh, MK.. Automatic Detection and Classification of Solitary Pulmonary Nodules from Lung CT Images. *IEEE*. 2014;294-299.
- [8] Dandil, E.; Cakiroglu, M.; Eksi, Z.; Ozkan, M.; Kurt, O.; Canan, A. “Artificial neural network-based classification system for lung nodules on computed tomography scans”. *Soft Computing and Pattern Recognition (SoCPaR)*. 2014. 382-386.
- [9] Han, F. Guopeng Z; Huafeng W; Bowen S; Hongbing L; Dazhe Z; Hong Z ; Zhengrong L, et al. A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database. *IEEE* 2013; 14-18.
- [10] Breiman, L. Random forests. *Machine learning*. 2001;1:5-32.
- [11] HAYKIN, Simon S. et al. *Neural networks and learning machines*. Upper Saddle River, NJ, USA:: Pearson, 2009.
- [12] Carvalho Filho AO. de. Detecção automática de nódulos pulmonares solitários usando quality threshold clustering e MVR [Dissertação]. São Luís: Universidade Federal do Maranhão; 2013.

- [13] Armato III SG., McLennan G., Bidaut L., McNitt-Gray M.F., Meyer C.R., Reeves A.P., Zhao B., Aberle D.R., Henschke C.I., Hoffman E.A., et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys* 2011; 38:915-31.
- [14] CANCER IMAGING ARCHIVE. LIDC-IDRI. 2016 Fev. Disponível em: <https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI>
- [15] Haralick, R. M.; Shanmugam, K.; Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1976; 3(6):610-21.
- [16] Magurran, AE. Measuring biological diversity. *African Journal of Aquatic Science*. 2004; 29(2):285-86.
- [17] Sousa US. Classificação de massas na mama a partir de imagens mamografias usando o índice de diversidade de shannon-wiener [Dissertação]. São Luís: Universidade Federal do Maranhão; 2011.
- [18] Oliveira FSS. Classificação de tecidos da mama em massa e não-massa usando índice de diversidade taxonômico e máquina de vetores de suporte [Dissertação]. São Luís: Universidade Federal do Maranhão; 2013.
- [19] Faith DP. Conservation evaluation and phylogenetic diversity. *Biol. Conserv*. 1992,61(1):1-10.
- [20] Mitchell TM. Machine learning. 1a ed. New York: McGraw Hill Higher Education; 1997.
- [21] Hall, M. Frank, E. Holmes, G. Pfahringer, B. Reutemann, P. Witten, I. H. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, v. 11, n. 1, p. 10-18, 2009.
- [22] Landis JR, Gary GK. The measurement of observer agreement for categorical data. *Biometrics*;1977. 159-74.
- [23] Martinez EZ; Louzada NF; Pereira BB. A curva ROC para testes diagnósticos. *Cadernos Saúde Coletiva*;2003. 7-31.
- [24] Metz CE.ROC methodology in radiologic imaging. *Invest. Radiol*;196. 21(9): 720-33.

Contato

Otília de Sousa Santos¹

E-mail: otylyasousa@gmail.com

Telefone: (89) 99436 -2647

Thayane de O. Simões¹

E-mail: thayanesimoes19@gmail.com

Telefone:(89) 9911-0323

Laércio N. Mesquita¹

E-mail: laerciomesquita90@gmail.com

Telefone:(86) 9918 –6191

Antonio O. de Carvalho Filho¹

E-mail:antoniooseas@gmail.com

Telefone:(89) 99465–9693

Alcilene D. de Sousa¹

E-mail:alcileneluzsousa@gail.com

Telefone:(89) 99422–1033

Endereço¹: Campus Senador Helvídio

Nunes de Barros -Universidade Federal do

Piauí (UFPI), Av. Cícero Eduardo -Junco,

CEP64607675, Picos, PI, Brasil

