Otília de Sousa Santos

Análise Filogenética Para Diferenciação entre Nódulos Malignos e Benignos

Picos - PI
Junho de 2017

Otília de Sousa Santos

Análise Filogenética Para Diferenciação entre Nódulos Malignos e Benignos

Trabalho de Conclusão de Curso submetido à Coordenação do Curso de Bacharelado em Sistemas de Informação da Universidade Federal do Piauí, Campus Senador Helvídio Nunes de Barros, no período 2017.1 como requisito para obtenção do título de Bacharel em Sistemas de Informação. Orientadora: Prof. Msc. Alcilene Dalília de Sousa.

Universidade Federal do Piauí
Campus Senador Helvídio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
Junho de 2017

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

S237a Santos, Otília de Sousa.
Análise filogenética para diferenciação entre nódulos malignos e benignos / Otília de Sousa Santos– 2017.
CD-ROM : il.; 4 ¾ pol. (26 f.)
Trabalho de Conclusão de Curso (Curso Bacharelado em Sistemas de Informação) – Universidade Federal do Piauí, Picos, 2017.
Orientador(A): Prof.ª Ma. Alcilene Dalília de Sousa

1. Imagens Médicas. 2.Diagnóstico Pulmonar. 3.Índices de Diversidade Filogenética. I. Título.

CDD 005.118

ANÁLISE FILOGENÉTICA PARA DIFERENCIAÇÃO ENTRE NÓDULOS MALIGNOS E BENIGNOS

OTÍLIA DE SOUSA SANTOS

Monografia aprovada como exigência parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Data de Aprovação

Picos – PI, 19 de junho de 2013

Aleilene Dalilia de Sousa
Profª. Ma. Aleilene Dalilia de Sousa
Orientador

Prof. Dr. Antonio Oseas de Carvalho Filho
Membro

Profª. Ma. Patricia Medyna Lauritzen de Lucena Drumond
Membro

Agradecimentos

Uma batalha termina. Uma etapa da minha vida foi concluída e olhando esse momento, vejo tudo o que passei, estudei e lutei me prepararam e me fortaleceram para as futura batalhas que me aguardam. Agradeço a Deus que me iluminou em todos os caminhos. A minha Orientadora Alcilene pelo apoio e dedicação e meus cooreintadores Antonio Oseas e Patricia Medyna. Aos meus pais Raimunda e Alzir, pelo amor, dedicação e confiança. As minhas irmãs Alzira, Alzirene e Almira pelo apoio e carinho, ao meu companheiro e amigo Hélio pelo amor, paciência, companherismo e ajuda todos esses anos. A minha filha Valentina por tornar essa batalha mais alegre e me transformar para melhor. A minha parceira de trabalho e amiga Thayane Simões pois tudo que conseguimos nesse trabalho foi fruto de muita dedicação e esforço e obrigada pela parceria ao desenvolver o trabalho somos uma dupla imbatível, a minha amiga Deyse Thaina por sempre me ajudar e está ao meu lado em todos os momentos. À minha turma em especial Ramon, Fernando, Angra, Laise, Anderson, Cidronio, Domingas, Walisson e a todos os meus amigos, pelo apoio e disposição em me ajudar todas as vezes que precisei. Aos parceiros da ICV pela momentos de aprendizados e descontração que vivemos, pelos sorriso, choros, pelas pizzas, por tudo.Aos professores pelos ensinamento e conselhos. Vocês foram imprescindíveis. Meu muito obrigada!

Que teu coração deposite toda a sua confiança no Senhor! Não te firmes em tua própria sabedoria! Sejam quais forem os teus caminhos, pensa nele, e ele aplainará tuas sendas.

Provérbios 3,5-6

Resumo

No fim do século XX, o câncer de pulmão tornou-se uma das principais causas de morte inevitáveis. Este câncer tem sido o mais comum entre os tumores malignos. Em virtude disso, há necessidade de diagnóstico precoce para que possa ser empregado o tratamento eficaz e, com isso, colaborar para uma maior taxa de sobrevida. Este trabalho, apresenta uma metodologia de auxílio no diagnóstico por computador (*Computer-Aided Diagnosis* -CADx) para a classificação de malignidade ou benignidade dos nódulos pulmonares em imagens de tomografia computadorizada, auxiliando os especialistas no diagnóstico e tratamento deste tipo de câncer. Os índices de diversidade filogenética foram utilizados 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.

Palavras-chaves: Imagens Médicas, Diagnóstico Pulmonar, Índices de Diversidade Filogenética.

Abstract

At the end of the 20th century, lung cancer became one of the main causes of inevitable death. This cancer has been the most common among malignant tumors. By virtue of this, there is a need for early diagnosis so that the effective treatment can be employed and with this, collaborating for a higher rate of life. This work presents a methodology of aid in computer diagnosis Computer-Aided diagnosis-CADx for the classification of malignancy or kindness of the pulmonary nodules in computed tomography images, assisting specialists in the diagnosis and treatment of this type of cancer. The use of the indexes of phylogenetic diversity to extract the characteristics of the nodules, the classification is accomplished with the tool Weka using multiple classifiers, validation of the results with the metrics kappa, Area under the curve, sensitivity, specificity and accuracy. The tests showed well-purpose and robust results for a CADx methodology with a accuracy of 98.1 %, sensitivity 98.7 %, specificity 97.9 %, a kappa of 0.95 and an area under the curve of 0.99. The results obtained proved the good performance of texture-characteristics extraction techniques through the indexes presented.

Keywords: Tomography, Diagnosis, Diversity Index.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

INCA	Instituto Nacional de Cancer
PDI	Processamento Digital de Imagens
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnosis*
TC	Tomografias Computadorizadas
ELCAP	*Early Lung Cancer Action Program*
LIDC-IRDI	*Lung Image Database* Consortium
PD	*Phylogenetic Diversity*
SPD	*Sum of Phylogenetic Distances*
MNND	*Mean Nearest Neighbor Distance*
PSV	*Phylogenetic Species Variability*
PSR	*Phylogenetic Species Richness*
RF	*Random Forest*
RT	*Random Tree*
MLP	*MultiLayerPerceptron*
UH	*Unidades Hounsfield*
XML	*eXtensible Markup Language*
WEKA	*Waikato Environment for Knowledge Analysis*
AUC	*Area Under the Curve*
S	Sensibilidade
E	Especificidade
A	Acurácia
K	índice Kappa
VP	Verdadeiro Positivo

FP	Falso Positivo
VN	Verdadeiro Negativo
FN	Falso Negativo

Sumário

1 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 (INCA, 2016).

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 câncer (INCA, 2016). 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 o 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 Sousa (2011), 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, sum of phylogenetic distances, phylogenetic species variability, phylogenetic species richness* e *Mean nearest neighbor distance*, e b) uso de árvores filogenéticas para caracterização dos nódulos pulmonares.

1.1 Objetivo Geral

Desenvolver uma metodologia automática para classificação de nódulos pulmonares baseada na análise de textura, extraída com os índices de diversidade filogenéticos com intuito de identificar a natureza dos nódulos pulmonares em maligno ou benigno, sendo assim, uma segunda opinião para o especialista da área.

1.2 Organização do Trabalho

Este trabalho está organizado em 6 (seis) capítulos. No capítulo 2 (dois) encontra-se os trabalhos relacionados. No capítulo 3 (três), faz-se a apresentação da metodologia utilizada. No capítulo 4 (quatro) apresenta-se os resultados obtidos, contendo testes realizados e trabalhos futuros. No capítulo 5 (cinco) a Conclusão e no capítulo 6 (seis)as Publicações.

2 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.

Fangfang Guopeng Zhang (2013) apresentaram um trabalho que avaliou o desempenho de duas dimensões (2D) e características de textura de três dimesões(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 nas base de características de textura 3D de Haralick.

Dandil E. (2014) 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.

Mukherjee Amlan Chakrabarti (2014), 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.

Froner (2015), 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 Osslan Osiris Vergara Villegas (2015), 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%.

Os trabalhos relacionados expostos nesta seção, 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.

3 Metodologia Proposta

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) (Helmus Thomas J. Bland (2007)); para a classificação utilizou-se *Random Forest* (Breiman (2001)), *Random Tree*, *MultiLayerPerceptron* (Haykin (2008)) e JRip e, por fim, a validação dos resultados. Figura 1 apresenta um resumo das etapas seguidas pela metodologia proposta.

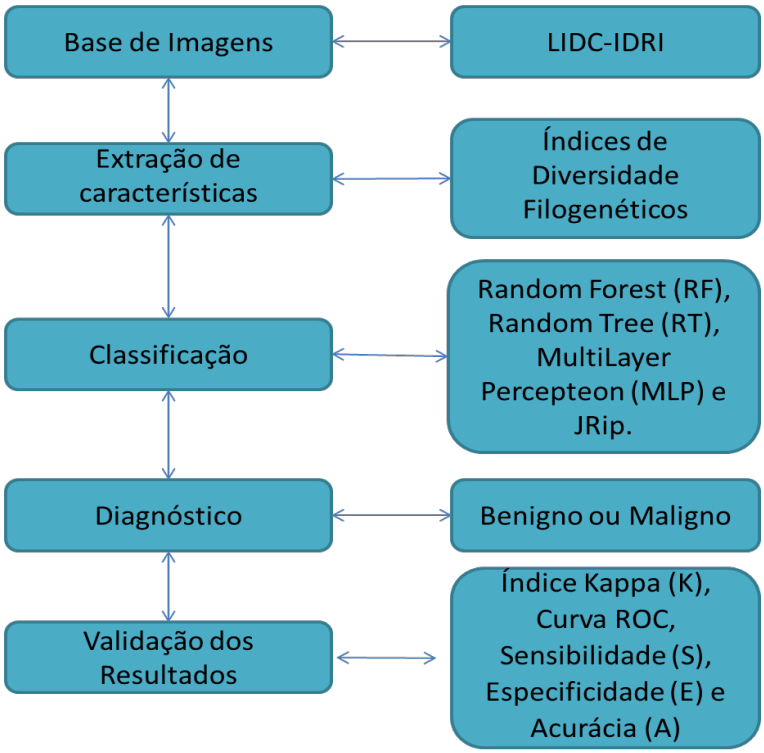


Figura 1 – Metodologia proposta para diferenciação entre nódulos pulmonares malignos e benignos.

3.1 Base de Imagens

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 maligno (CarvalhoFilho (2013)) , como demonstra a Figura 2.

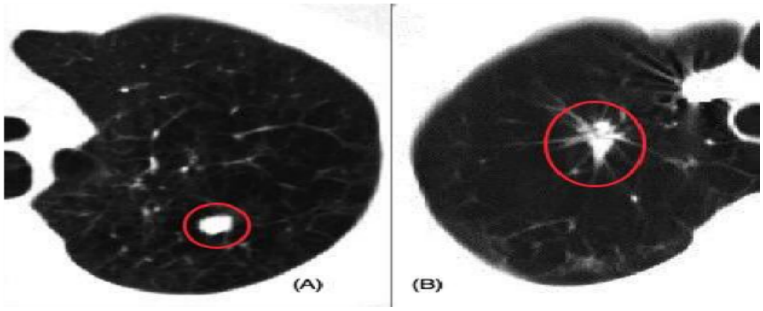


Figura 2 – Exemplos de TC, onde é possível verificar o aparecimento de nódulos pulmonares: (A) TC com o nódulo benigno e (B) TC com nódulo maligno (Armato et al. (2011)).

No desenvolvimento desse trabalho utilizou-se a base de imagens LIDC-IDRI Armato et al. (2011), 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 (*itálico eXtensible Markup Language*) contendo as marcações e as avaliações de 4 especialistas (LIDC-IDRI, 2016).

3.2 Índices de Diversidade Filogética

A definição de textura encontrada na literatura é descrita de diversas formas. Segundo Haralick, Shanmugam e Dinstein (1976), 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 foi utilizada como base para a extração das características dos nódulos pulmonares. Para Magurran (2004) 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 mesma, como apresentado na Figura 3.

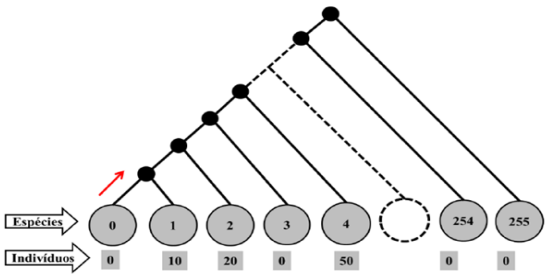


Figura 3 – Árvore enraizada na forma de cladograma inclinado Fonte: (Oliveira (2013))

Utilizou-se 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 UH 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 Faith (1992). Quanto mais ramos a árvore filogenética possuir, mais distintos serão os grupos taxonômicos. 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 3.1

$$PD = \frac{B * \sum_i^s LiAi}{\sum_i^s LiAi} \tag{3.1}$$

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

Para Webb (2000), 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 3.2,

$$MNND = \sum_{i}^{s} min\,(dmn)\,am \tag{3.2}$$

onde S é número de espécies do conjunto focal, m e n são espécies, dmn é a distância filogenética entre m e n e am é a abundância das espécies de m.

Segundo Helmus Thomas J. Bland (2007), 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 3.3,

$$SPD = \frac{\frac{S(S-1)}{2} * \sum\sum^{m} < n^{d}mn^{a}m^{a}n}{sum\sum^{m} < n^{d}mn^{a}m^{a}n} \tag{3.3}$$

onde S é número de espécies do conjunto focal, m e n são espécies, dmn é a distância filogenética entre m e n, am é a abundância das espécies de m e an é a abundância das espécies de n.

No trabalho desenvolvido por Helmus Thomas J. Bland (2007), 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 3.4 define a variabilidade de espécies filogenéticas, que resume o grau em que as espécies em uma comunidade são filogeneticamente relacionadas,

$$PSV = \frac{ntrC - \sum c}{n\,(n-1)} = 1 - \overline{C} \tag{3.4}$$

onde representa-se a soma dos valores da diagonal de uma matriz C, o somatorio de todos os valores da matriz, n é o número de espécies e c 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 3.5, onde o valor do PSR é a multiplicação do número espécies n pela variabilidade da comunidade.

$$PSR = n * PSV \tag{3.5}$$

3.3 Classificação e Validação

Segundo Mitchell (1997), 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 Hall Eibe Frank e Witten (2009).

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 é 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.

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) (Landis e Koch (1977)), Sensibilidade (S), Especificidade (E) (Martinez, Louzasa-Neo e Pereira (2003)), Acurácia (A) (Metz (1986)) e índice Kappa (K).

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, Louzasa-Neo e Pereira (2003) 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 Metz (1986), 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 Koch (1977), 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.

4 Resultados

Para os testes realizados neste trabalho, utilizou-se a base LIDC-IDRI com 1402 nódulos, sendo 1009 benignos e 393 malignos. As características foram extraídas a partir dos índices de diversidade filogenética. Os classificadores escolhidos foram *Random Forest* (RF), *Random Tree* (RT), *MultiLayerPerceptron* (MLP) e JRip, pois são os mais utilizados na literatura, 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 et al.(17)	ELCAP and LIDC	90,9%	73,91%	82%
Dandil et al.(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.

5 Conclusão

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 sobrevida 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 aprendizado de máquina, 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.

6 Publicações

SANTOS, O. S; SIMÕES. T. O.; SOUSA, A. D.; CARVALHO FILHO, A. O; DRUMOND, P. M. L. L. Diferenciação dos padrões de malignidade e benignidade em imagens de tomografia computadorizada usando índice de diversidade filogenético e SVM. Escola Regional de Informática do Piauí – ERIPI,2016.

SANTOS, O. S; SIMÕES. T. O.; MESQUITA, L. N.; SOUSA, A. D.; CARVALHO FILHO, A. O. Análise filogenética para diferenciação entre nódulos malignos e benignos. Sociedade Brasileira de Informática em Saúde – SBIS, 2016.

SIMÕES. T. O.; **SANTOS, O. S**; CARVALHO FILHO, A. O; DRUMOND, P. M. L. L.; SOUSA, A. D. Redução de falsos positivos em imagens de tomografia computadorizada, usando índice de diversidade filogenético e SVM. Escola Regional de Informática do Piauí – ERIPI,2016.

SIMÕES. T. O.; **SANTOS, O. S**; MESQUITA, L. N.; CARVALHO FILHO, A. O; SOUSA, A. D. Redução de falsos positivos baseado nas relações filogenéticas entre espécies. Sociedade Brasileira de Informática em Saúde – SBIS, 2016.

Referências

ARMATO, S. G. 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*, v. 38, n. 2, p. 915–31, 2011. ISSN 0094-2405. Disponível em: <http://www-.biomedsearch.com/nih/Lung-Image-Database-Consortium-LIDC/21452728.html>. Citado 2 vezes nas páginas 8 e 17.

BREIMAN, L. Random forests. machine learning. v. 45, n. 1, p. 5–32, 2001. Citado na página 16.

CARVALHOFILHO, A. O. *Detecção automática de nódulos pulmonares solitários usando quality threshold clustering e MVR.* Dissertação (Dissertação de Mestrado), 2013. Citado na página 17.

DANDIL E., C. M. E. Z. O. M. K. O. C. A. Artificial neural network-based classification system for lung nodules on computed tomography scans. In: *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of.* [S.l.: s.n.], 2014. p. 382–386. Citado na página 15.

FAITH, D. P. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, v. 61, n. 1, p. 1 – 10, 1992. ISSN 0006-3207. Disponível em: <http:/-/www.sciencedirect.com/science/article/pii/0006320792912013>. Citado na página 18.

FANGFANG GUOPENG ZHANG, H. W. B. S. H. L. D. Z. H. Z. a. L. H. A texture feature analysis for diagnosis of pulmonary nodules using lidc-idri database. IEEE, p. 14–18, 2013. Citado na página 15.

FRONER, A. P. P. *Caracterizaçãoo de nódulos pulmonares em imagens de tomografia computadorizada para fins de auxílio ao diagnóstico.* Dissertação (Dissertação de Mestrado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2015. Citado na página 15.

HALL EIBE FRANK, G. H. B. P. P. R. M.; WITTEN, I. H. *The WEKA Data Mining Software: An Update.* 2009. 10–18 p. ACM SIGKDD explorations newsletter. Disponível em: <www.cms.waikato.ac.nz/˜ml/publications/2009/weka_update.pdf>. Citado na página 20.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, IEEE, n. 6, p. 610–621, 1976. Citado na página 17.

HAYKIN, S. *Neural networks and learning machines.* New Jersey: Bookman, 2008. Citado na página 16.

HELMUS THOMAS J. BLAND, C. K. W. A. R. I. M. R. Phylogenetic measures of biodiversity. *the american naturalist*, v. 169, n. 3, p. 1–16, 2007. Citado 2 vezes nas páginas 16 e 19.

INCA. Estimativas 2016: Incidência de Câncer no Brasil. 2016. Disponível em: <http://www.inca.gov.br/estimativa/2016/estimativa-2016-v11.pdf/>. Citado na página 13.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159–174, 1977. Disponível em: <http://www.jstor.org-/stable/2529310 .> Citado na página 20.

LIDC-IDRI. 2016. Disponível em: <hhttps://wiki.cancerimagingarchive.net/display-/Public/LIDC-IDRI>. Citado na página 17.

MAGURRAN, A. E. Measuring biological diversity. *African Journal of Aquatic Science*, v. 29, n. 2, p. 285–286, 2004. Citado na página 17.

MARTINEZ, E. Z.; LOUZASA-NEO, F.; PEREIRA, B. B. de. A curva roc para testes diagnosticos. *Cad Saúde Coletiva*, v. 11, n. 1, p. 7–31, 2003. Disponível em: <www.po.ufrj.br/.../2003_a_curva_ROC_para_testes_diagnosticos_cadernos_saude_co>. Citado na página 20.

METZ, C. E. Roc methodology in radiologic imaging. *Investigative radiology*, v. 21, n. 9, p. 720–733, 1986. Citado na página 20.

MITCHELL, T. *Machine Learning*. New York: McGraw Hill Higher Education: Bookman, 1997. Citado na página 19.

MUKHERJEE AMLAN CHAKRABARTI, S. H. S. M. K. J. Automatic detection and classification of solitary pulmonary nodules from lung ct images. *Systems, Man and Cybernetics, IEEE Transactions on*, IEEE, p. 294–299, 2014. Citado na página 15.

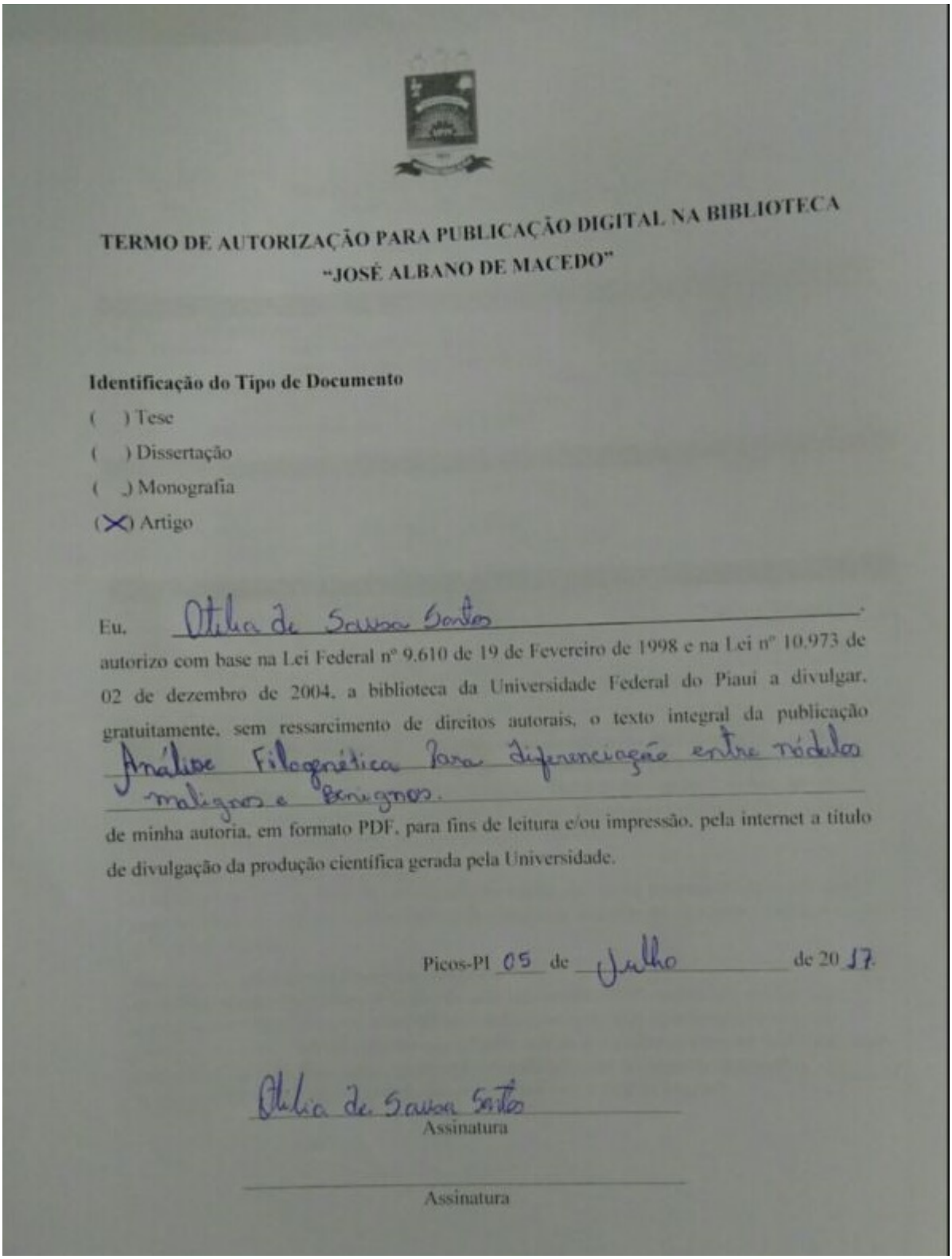
OLIVEIRA, F. S. S. de. *Classificacao de Tecidos da Mama em Massa e Nao–Massa usando indice de Diversidade Taxonomico e Maquina de Vetores de Suporte.* Dissertação (Dissertação de Mestrado), 2013. Citado 2 vezes nas páginas 8 e 18.

OROZCO OSSLAN OSIRIS VERGARA VILLEGAS, V. G. C. S. H. d. J. O. D. M. d. J. N. A. H. M. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine. *BioMedical Engineering OnLine*, v. 14, p. 9, 2015. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4329222/>. Citado na página 15.

SOUSA, U. S. *Classificação De Massas na Mama a partir De Imagens Mamográficas Usando índice De Diversidade De Shannon-Wiener.* Dissertação (Dissertação de Mestrado) — Universidade Federal do Maranhão, 2011. Citado na página 13.

WEBB, C. O. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *the american naturalist*, v. 156, n. 2, p. 145–155, 2000. Citado na página 19.

A Termo de Autorização



TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA "JOSÉ ALBANO DE MACEDO"

Identificação do Tipo de Documento

( ) Tese
( ) Dissertação
( ) Monografia
(X) Artigo

Eu, Otilia de Sousa Santos, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação Análise Filogenética para diferenciação entre tóxicos malignos e benignos, de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI 05 de Julho de 2017.

Otilia de Sousa Santos
Assinatura

Assinatura