Robherson Wector de Sousa Costa

Classificação de nódulos pulmonares em maligno e benigno usando o índice de diversidade taxonômica e distância média filogenética

São Luís - MA

2019

Robherson Wector de Sousa Costa

Classificação de nódulos pulmonares em maligno e benigno usando o índice de diversidade taxonômica e distância média filogenética

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Aristófanes Corrêa Silva

São Luís - MA

2019

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Costa, Robherson Wector de Sousa.
Classificação de nódulos pulmonares em maligno e benigno usando o índice de diversidade taxonômica e distância média filogenética / Robherson Wector de Sousa Costa. - 2020.
51 f.

Orientador(a): Aristófanes Corrêa Silva.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luis - MA, 2020.

1. Árvores filogenéticas. 2. Imagens médicas. 3. Índice de distância média filogenética. 4. Índice de diversidade taxonômica. 5. Nódulo pulmonar. I. Silva, Aristófanes Corrêa. II. Título.

Robherson Wector de Sousa Costa

Classificação de nódulos pulmonares em maligno e benigno usando o índice de diversidade taxonômica e distância média filogenética

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho aprovado em 07 de Janeiro de 2020

Prof. Dr. Aristófanes Corrêa Silva
Orientador
Universidade Federal do Maranhão

Prof[a]. Dra. Simara Vieira da Rocha
Examinador
Universidade Federal do Maranhão

Prof. Dr. Giovanni Lucca França da Silva
Examinador
Universidade Federal do Maranhão

São Luís - MA
2019

Agradecimentos

Agradeço a Deus em primeiro lugar, por me proporcionar saúde e forças para enfrentar a difícil batalha que é uma graduação. Ao Deus que sempre esteve comigo e me deu coragem e força em todos os momentos em que pensei em desistir por causa de problemas ou dificuldades enfrentadas durante a graduação.

À minha família, em especial ao meu finado avô Antônio José que me mostrou que a educação é essencial no crescimento de um ser humano, e à minha avó Raimunda que estive comigo em todos os momentos da minha vida me apoiando e incentivando de varias maneiras. Aos meus pais que sempre me apoiaram na minha caminhada, conversando e aconselhando. E aos demais parentes que contribuíram de forma direita ou indireta na minha formação.

À minha noiva Jéssica, que esteve ao meu lado nesses últimos meses de curso, me apoiando e incentivando sempre.

Em especial ao professor Aristófanes Corrêa Silva, meu orientador (e grande amigo) que esteve comigo a maior parte do curso, e com isso contribuiu para meu crescimento em geral, me provando que sempre posso melhorar e me mostrando que o meu limite sou eu mesmo que defino.

Ao Núcleo de Computação Aplicada (NCA), pelas oportunidades de crescimento através das pesquisas e desenvolvimentos que tive o prazer de participar e que tiveram uma contribuição enorme em minha vida. E por me proporcionar o ambiente, profissionais e equipamentos necessários para o desenvolvimento deste trabalho. E por me proporcionar diversos amigos que contribuíram de forma significativa para minha formação.

À todos o professores da UFMA que contribuíram para me desenvolvimento tanto em sala de aula quanto fora dela. Aos professores Anselmo Cardoso de Paiva, Geraldo Braz Júnior e Simara Vieira da Rocha que como orientadores e chefes me orientaram e aconselharam durante essa caminhada.

À Universidade Federal do Maranhão(UFMA), por proporcionar o meu crescimento tanto pessoal como profissional através de sua estrutura e de seus excelentes profissionais.

E a todos que contribuíram comigo de forma direta ou indireta.

"Consagre ao Senhor tudo o que você faz,
e os seus planos serão bem-sucedidos."

(Provérbios, 16:3)

Resumo

O câncer de pulmão apresenta a maior causa de morte entre os pacientes em todo o mundo, além de ser uma das menores taxas de sobrevida após o diagnóstico. Portanto, existe uma necessidade crescente de utilizar meios alternativos de diagnóstico para esse tipo de tumor, tornando-se uma ferramenta importante, pois reduz o grau de incerteza no diagnóstico, fornecendo ao médico especialista uma fonte de informação adicional. A metodologia proposta é baseada em técnicas de processamento de imagem e reconhecimento de padrões. Utilizamos índices da Ecologia, chamados índice de diversidade taxonômica (Δ) e distância média filogenética (MPD), como descritores de textura, para sugerir uma malignidade ou benignidade do nódulo pulmonar. O cálculo desses índices é baseado em árvores filogenéticas. Esses descritores são usados para o algoritmo genético que gera o melhor modelo de treinamento para ser usado com um classificador de máquina de vetor de suporte. Nos testes, foram usados 1.405 nódulos (1.011 benignos e 394 malignos) do banco de dados LIDC-IDRI. A base foi dividida em dois grupos: treino e teste, com proporções de 80% e 20%, respectivamente. A metodologia apresenta sensibilidade de 93,42%, especificidade de 91,21%, precisão de 91,81% e uma curva ROC de 0,94.

Palavras-chave: Imagens médicas, nódulo pulmonar; árvores filogenéticas; índice de diversidade taxonômica; índice de distância média filogenética.

Abstract

Lung cancer is the leading cause of death among patients worldwide, and is one of the lowest survival rates after diagnosis. Therefore, there is a growing need to use alternative diagnostic means for this type of tumor, making it an important tool as it reduces the degree of uncertainty in the diagnosis by providing the specialist with an additional source of information. The proposed methodology is based on image processing and pattern recognition techniques. We used Ecology indices, called taxonomic diversity index (Δ) and phylogenetic mean distance (MPD), as texture descriptors to suggest malignancy or pulmonary node initiation. The calculation of these indices is based on phylogenetic trees. These descriptors are used for the genetic algorithm that generates the best training model for use with a support vector machine classifier. In the tests, 1,405 nodules (1,011 benign and 394 malignant) from the LIDC-IDRI database were used. The base was divided into two groups: training and test, with proportions of 80% and 20%, respectively. A methodology that presents sensitivity of 93.42%, specificity of 91.21%, precision of 91.81% and a ROC curve of 0.94.

Keywords: Medical images, pulmonary nodule; phylogenetic trees; taxonomic diversity index; phylogenetic mean distance index.

Lista de ilustrações

Figura 1 – Exemplo de nódulo pulmonar 19
Figura 2 – Exemplo de uma imagem que possui três níveis de cinza (espécies) que é preto, o cinza e o branco. A quantidade de pixels (indivíduos) de preto é 4, de cinza é 3 e de branco é 2. 22
Figura 3 – Representação de uma árvore filogenética para alguns primatas. 23
Figura 4 – Exemplo representando uma imagem em uma árvore taxonômica (à esquerda) e sua matriz de distância (à direita). 24
Figura 5 – Árvore filogenética enraizada na forma de cladograma inclinado. 24
Figura 6 – Separação entre duas classes através de hiperplanos. 26
Figura 7 – Metodologia Proposta. 29
Figura 8 – Ilustração do resumo das marcações dos nódulos. 30
Figura 9 – Abordagem de máscara interna. 32
Figura 10 – Abordagem de máscara externa. 33
Figura 11 – Árvore 1: árvore enraizada na forma de cladograma inclinado. 34
Figura 12 – Relações entre as especies(UH). 35
Figura 13 – Árvore 2: modelo criado a partir da árvore 1, excluindo as espécies que não possuem indivíduos. 36
Figura 14 – Árvore 3: modelo criado a partir da árvore 1, modificando as arestas . 37
Figura 15 – Teste 1: Modelo de treino desbalanceado 40
Figura 16 – Teste 2: Modelo de treino balanceado 40

Lista de tabelas

Tabela 1 – Correspondência entre termos da Biologia e nosso trabalho 33
Tabela 2 – Resultado obtido pelo MVS sem balanceamento 41
Tabela 3 – Melhor resultado obtido pelo AG e MVS no conjunto de validação . . . 41
Tabela 4 – Resultados obtidos com nossa metodologia 41
Tabela 5 – Comparação dos resultados do nosso trabalho com os trabalhos relacionados. 42
Tabela 6 – Comparação dos resultados do nosso trabalho com os trabalhos relacionados. 43

Lista de abreviaturas e siglas

AG	Algoritmo Genético
ANN	Artificial Neural Network
CAD	Detecção Assistida por Computador
CADx	Diagnóstico Assistido por Computador
Δ	Índice de Diversidade Taxonômica]
DDSM	Digital Database for Screening Mammography
FN	Falso Negativo
FP	Falso Positivo
GLCM	GrayLevel Co-occurrence Matrix
LIDC-IDRI	Lung Image Database Consortium image collection
MPD	Distância Média Filogenética
MVS	Maquina de Vetores de Suporte
RBF	Radial Basis Function Neural Network
ROI	Region Of Interest
SOM	Self-Organizing Maps
TC	Tomografia Computadorizada
UH	Unidade de Hounsfield

Sumário

1 INTRODUÇÃO . . . 13
1.1 Objetivo . . . 14
1.1.1 Objetivos Específicos . . . 14
1.2 Trabalhos relacionados . . . 14

2 FUNDAMENTAÇÃO TEÓRICA . . . 19
2.1 Nódulo Pulmonar Solitário . . . 19
2.2 Tomografia Computadorizada . . . 20
2.3 Quantização Uniforme . . . 20
2.4 Análise de Textura . . . 21
2.4.1 Índice de Diversidade . . . 21
2.4.2 Diversidade Filogenética . . . 22
2.4.3 Índices Taxonômicos . . . 24
2.4.4 Distância Filogenética Média (Mean phylogenetic distance- MPD) . . . 25
2.5 Máquina de Vetores de Suporte . . . 25
2.6 Algoritmo Genético . . . 26
2.7 Validação dos Resultados . . . 27

3 METODOLOGIA . . . 29
3.1 Aquisição de Imagens . . . 29
3.2 Segmentação dos Nódulos . . . 31
3.3 Extração de Características . . . 31
3.3.1 Abordagem de máscara interna e externa . . . 32
3.3.2 Árvores filogenéticas . . . 33
3.3.2.1 Árvore enraizada na forma de cladograma inclinado . . . 34
3.3.2.2 Árvore enraizada na forma de cladograma inclinado excluindo as espécies sem indivíduos . . . 35
3.3.2.3 Árvore enraizada na forma de cladograma inclinado modificando as arestas . . . 36
3.4 Treinamento e Validação . . . 36
3.5 Validação dos Resultados . . . 38

4 RESULTADOS . . . 39
4.1 Aquisição de Imagens . . . 39
4.2 Extração de Características . . . 39
4.3 Treino e Validação . . . 39
4.4 Classificação . . . 40

4.4.1 Comparação com trabalhos relacionados 42

5 CONCLUSÕES 45
5.1 Trabalhos Futuros 45

REFERÊNCIAS 47

1 Introdução

O câncer de pulmão é o segundo mais comum em homens e mulheres no Brasil (sem contar o câncer de pele não melanoma). É o primeiro em todo o mundo desde 1985, tanto em incidência quanto em mortalidade. Cerca de 13% de todos os casos novos de câncer são de pulmão. A última estimativa mundial (2012) apontou incidência de 1,8 milhão de casos novos, sendo 1,24 milhão em homens e 583 mil em mulheres.(INCA, 2019)

A taxa de incidência vem diminuindo desde meados da década de 1980 entre homens e desde meados dos anos 2000 entre as mulheres. Essa diferença deve-se aos padrões de adesão e cessação do tabagismo constatados nos diferentes sexos. No Brasil, a doença foi responsável por 26.498 mortes em 2015. No fim do século XX, o câncer de pulmão se tornou uma das principais causas de morte evitáveis. (INCA, 2019)

Um nódulo pulmonar é caracterizado por ser uma opacidade arredondada no pulmão com até 3 cm de diâmetro, envolta por parênquima pulmonar (HANSELL et al., 2008). Lesões pulmonares maiores que 3 cm de diâmetro, são consideradas massas e muitas vezes malignas (FUJIMOTO; WISTUBA, 2014; GOULD et al., 2001; HANSELL et al., 2008). O seu diagnóstico e tratamento precoce aumentam as chances de sobrevivência do paciente em cerca de 90% (LEDERLIN et al., 2013). Nesse sentido, os exames por imagem demonstram ser boas alternativas para detecção e diagnóstico de nódulos pulmonares, dentre eles, destacam-se os exames de tomografia computadorizada (TC) (SRICHAI; NAIDICH; AL., 2007). A detecção do nódulo usando TC não é uma tarefa simples, pois os nódulos podem ter contrastes semelhantes à outras estruturas, baixa densidade, e um pequeno tamanho em uma área de anatomia complexa (pode ser ligado aos vasos sanguíneos ou nas fronteiras do pulmão), entre outras questões(LEEF; KLEIN, 2002).

Nas últimas décadas, tem surgido um grande interesse no desenvolvimento e uso de técnicas de processamento digital de imagens em TC, com o objetivo principal de aumentar a precisão do diagnóstico, proporcionando ao especialista uma segunda opinião. Essas técnicas foram usadas em conjunto para desenvolver sistemas de detecção assistida por computador (CAD) / diagnóstico assistido por computador (CADx).

Na maioria dos sistemas CADx, o estágio de extração de características é baseado em: 1) geometria, que mede, por exemplo, quão circular é o nódulo e 2) na textura, que descreve aspectos do nódulo com base em sua distribuição de níveis de cinza. Para nossa metodologia, usaremos apenas características de textura. O índice de diversidade taxonômica e a distância média filogenética foram utilizados para descrever a textura dos nódulos benignos e malignos. O primeiro considera a abundância das espécies e a relação taxonômica entre elas. Por outro lado, o segundo representa a distância taxonômica média entre

dois indivíduos de espécies distintas. Esses índices são baseados na distância filogenética, considerando a arquitetura de uma árvore enraizada na forma de um cladograma inclinado. Particularmente, este trabalho contribui para a área por utilizar apenas características de textura baseadas no índice de diversidade taxonômica e distância média filogenética para caracterização de nódulos benignos e malignos.

1.1 Objetivo

O objetivo desse trabalho é estudar, implementar e analisar o uso de índices de diversidade para verificar a possibilidade de se estabelecer diferenças entre padrões malignos e benignos em nódulos pulmonares usando tomografia computadorizada.

1.1.1 Objetivos Específicos

- Estudar a viabilidade de aplicar índices de diversidade em imagens de nódulos pulmonares;
- Estudar a viabilidade de aplicar algoritmo de aprendizado de máquina supervisionado em imagens de nódulos pulmonares;
- Estudar e aplicar a técnica de reconhecimento de padrões: máquinas de vetor de suporte, para testar as características produzidas no tocante ao poder de discriminação das classes malignas e benignas dos nódulos pulmonares.
- Estudar o Índice de Diversidade Taxonômica e a Distância Filogenética Média e suas aplicações no reconhecimento de padrões.

1.2 Trabalhos relacionados

Índices de diversidade tem sido explorados como descritores de textura em outros trabalhos. A seguir apresentaremos trabalhos relacionados presentes na literatura.

A metodologia proposta por (KUMAR; RAMESH; AL., 2011) propõe o desenvolvimento de um sistema CADx. A base de imagens LIDC-IDRI foi utilizada. A segmentação dos nódulos é feita usando os padrões Fuzzy C-Means (FCM), Fuzzy-Possibilistic C-Means e Fuzzy-Possibilistic C-Means (WFCM). Desvio padrão, contraste e entropia padrão são utilizados para extração de características. Por fim foi utilizado o classificador MVS. Foram obtidos resultados de 80,36% de acurácia, 76,47% de sensibilidade e 82,05% de especificidade.

Em (ZINOVEV et al., 2011), propõe-se uma maneira de prever a distribuição das opiniões dos radiologistas utilizando um algoritmo de classificação para rotulação múltipla,

baseado em árvores de decisão. A base utilizada foi a Lung Image Database Consortium (LIDC). Os resultados são validados usando uma técnica baseada em limiar de distância e através da área sob a curva *Receiver Operating Characteristic* (ROC). Os resultados alcançados demonstram que a efetividade do método proposto é de 74,1% e uma ROC curve de 0.69.

(FARAG et al., 2011) analisa a eficácia dos descritores que analisam a geometria comuns em visão computacional, para a redução de falsos positivos e classificação de nódulos pulmonares em TC. Os resultados alcançados sob a base de imagens ELCAP mostrou que os descritores fornecem 2% de melhorias na especificidade quando utilizado em conjunto com o classificador K-nearest neighbors.

(NASCIMENTO, 2012) propôs uma metodologia de caracterização de nódulos pulmonares em benigno ou maligno. A metodologia foi aplicada na base LIDC com 73 nódulos, sendo 26 malignos e 47 benignos. Os índices de diversidade de Shannon e Simpson foram utilizados como descritores de textura. As características geradas foram submetidas à etapa de seleção de características com a utilização de análise discriminante stepwise. Após esta etapa foi realizada a classificação usando o MVS, onde foram obtidas taxas de sensibilidade de 85.64%, especificidade de 97.89% e acurácia de 92.78%.

Já em (ELIZABETH et al., 2012) é proposto um CADx para diagnóstico de câncer de pulmão. Em primeiro lugar, a imagem de TC foi pré-processada através da segmentação do parênquima pulmonar de cada fatia usando Greedy Snake Algorithm. As ROIs foram, em seguida, extraídas a partir do parênquima pulmonar utilizando um algoritmo de crescimento de região. As ROIs extraídas foram marcadas como nódulos cancerosos ou não-cancerosos com o auxílio de um especialista humano e, em seguida, as características de forma e textura foram extraídas de cada ROI. As características extraídas e o rótulo do ROI foram usadas para treinar uma rede neural com RBF. A partir dos resultados experimentais, foi encontrada uma precisão de 94,44%.

No trabalho de (AL-ABSI et al., 2012) é proposto um sistema CADx para câncer de pulmão. Nas etapas de extração e seleção de características, diferentes funções de *Wavelets* foram aplicadas, a fim de encontrar o que produziu o melhor resultado de acurácia. A base *Japanese Society of Radiologic Technology* foi utilizada para testar o método proposto. A base de imagens conta com 154 regiões de nódulos (anormais) e 92 regiões de não-nódulo (normal). Níveis de acurácia de 96% para a classificação foram alcançados.

(OROZCO et al., 2012) propõe uma alternativa computacional para classificar nódulos pulmonares utilizando espectros do bidimensional Discrete Cosine Transform (2D-DCT) e o bidimensional Fast Fourier Transform (2D-FFT). O MVS foi utilizado como classificador. Após os experimentos, foram obtidas uma sensibilidade e especificidade de 96.15% e 52.17%, respectivamente. A acurácia total foi de 82.66%.

Em (KREWER; GEIGER; AL., 2013) medidas de textura e forma foram extraídos dos nódulos pulmonares selecionados da base de imagens LIDC. Vários classificadores incluindo Árvores de Decisão, vizinho mais próximo, e MVS foram utilizados para classificar nódulos pulmonares malignos e benignos. Uma acurácia total de 90,91%.

Em (OROZCO; VILLEGAS; AL., 2013) é apresentado uma metodologia baseada em TC para classificação de nódulos pulmonares. As regiões de interesse foram selecionadas manualmente. Para caracterização dos candidatos são extraídas medidas baseadas em textura a partir de seu histograma. Na etapa de classificação é usado uma MVS com *Radial Basis Function* (RBF). A metodologia foi validade em 75 exames. Atingindo resultados nas proporções de 10 falsos negativos (FN) e 2 falsos positivos (FP), sensibilidade e especificidade de 96,15% e 52,17% e uma acurácia de 82,66%.

O CAD desenvolvido por (FILHO et al., 2013) apresenta uma metodologia automática para detecção e classificação de nódulo pulmonar. Ela pode ser resumida em três grandes etapas. 1) extração e a reconstrução do parênquima pulmonar, em seguida, é aplicado um melhoramento para realçar suas estruturas. 2) os candidatos a nódulo são segmentados. E na última fase, são extraídas características de forma e textura, em seguida, classificados usando MVS. Os resultados alcançados apontam uma sensibilidade de 85,91%, uma especificidade de 97,70% e uma acurácia de 97,55%.

O trabalho de (DANDIL et al., 2014) apresentou um CADx para classificar nódulos benignos e malignos. Foram utilizadas 128 imagens de TC, obtidas a partir de 47 pacientes. Self-Organizing Maps (SOM) foi utilizada para segmentação dos nódulos, logo após foi utilizado o método GLCM (graylevel co-occurrence matrix) para extração de características. Por fim ANN (Artificial Neural Network) foi utilizado para classificação dos nódulos. Os resultados obtidos foram 90.63% de acurácia, 92.30% de sensibilidade e 89.47% de especificidade.

O trabalho desenvolvido por (AKRAM et al., 2015) tem como objetivo detectar e classificar nódulos pulmonares. Para isso são utilizados descritores de textura baseado em estatística e Máquina de Vetores de Suporte (MVS). Os nódulos candidatos são extraídos com base em anotações de especialistas. Após a segmentação, faz-se a extração das características dos candidatos a nódulo, com base em técnicas estatísticas, e por fim, é feita a classificação com MVS. A sensibilidade alcançada foi de 96,31%.

Já em (OLIVEIRA et al., 2015) é apresentada uma metodologia para classificação de regiões extraídas de mamografias digitais em massa e não-massa. O banco de imagens Digital Database for Screening Mammography (DDSM) é usado para teste. Na descrição da textura da região de interesse são utilizados os índices de diversidade taxonômica (Δ) e distinção taxonômica (Δ^*). Para classificação das regiões foi utilizada a support vector machine (MVS). A metodologia apresentou uma acurácia média de 99.67%.

O trabalho de (FARAG et al., 2017) apresenta um método utilizando Três recursos baseados em (i) filtro Gabor, (ii) recursos de textura de padrão binário local de várias resoluções (LBP) e (iii) distância assinada fundida com LBP, que geram características combinatórias de forma e textura, são utilizados para fornecer descritores de recursos de elementos malignos e nódulos benignos e regiões não-nódulos de interesse. Os classificadores de máquinas de vetor de suporte (MVSs) e k-vizinho mais próximo (kNN) em estruturas em cascata serial e de duas camadas são otimizados e analisados para obter os melhores resultados de classificação dos nódulos. Um total de 1191 amostras de nódulos e não nódulos do banco de dados do Lung Image Data Consortium é usado para análise. A classificação usando classificadores MVS e kNN é examinada. Os resultados da classificação do MVS em cascata de duas camadas usando os recursos de Gabor mostraram resultados melhores em geral para identificar nódulos não-nódulos malignos e benignos com área média sob as curvas de características operacionais do receptor (AUC-ROC) de 0,99 e escore f1 médio de 0,975 sobre os dois níveis.

Já em (NIBALI; HE; WOLLERSHEIM, 2017) foi avaliada a eficácia de redes neurais convolucionais muito profundas na tarefa de classificação de malignidade de nódulos pulmonares em nível de especialista. Usando a arquitetura ResNet de última geração como base, exploramos o efeito do aprendizado do currículo, transfere o aprendizado e a profundidade variável da rede na precisão da classificação de malignidade. Foram utilizados 831 exemplos de nódulos do conjunto de dados LIDC/IDRI. Foram obtidos os resultados de 91.07% de sensibilidade, 88.64% de especificidade, acurácia de 89.35% e uma Curva ROC de 0,9459.

Em (MASOOD et al., 2018) foi proposto um Sistema de Apoio à Decisão Assistido por Computador em Câncer Pulmonar, usando o novo modelo de aprendizado profundo e informações sobre metástases obtidas do MBAN (Medical Body Area Network). O modelo proposto, DFCNet, baseia-se na rede neural profunda e convolucional profunda (FCNN), usada para classificar cada nódulo pulmonar detectado em quatro estágios do câncer de pulmão. O desempenho do trabalho proposto é avaliado em diferentes conjuntos de dados com diferentes condições de varredura. A comparação do classificador proposto é feita com as técnicas existentes da CNN. Utilizando o conjunto de dados LIDC-IDRI foram o btidos os resultados de 86.02% de acurácia, 83.91% de sensibilidade e 89.32% de especificidade.

Em (ZHAO et al., 2018) uma CNN híbrida do LeNet e AlexNet é construída através da combinação das configurações de camada do LeNet e dos parâmetros do AlexNet. Um conjunto de dados com 743 amostras de nódulos de imagem de TC é construído com base nas 1018 tomografias do LIDC para treinar e avaliar o modelo Agile CNN. Ao ajustar os parâmetros do tamanho do kernel, taxa de aprendizado e outros fatores, o efeito desses parâmetros no desempenho do modelo da CNN é investigado, e uma configuração otimizada da CNN é obtida finalmente. Foi obtida uma acurácia de 85.64%.

(LI et al., 2019) apresentam um algoritmo de fusão que combina recursos artesanais (IC) com os recursos aprendidos na camada de saída de uma rede neural convolucional profunda em 3D (CNN). Primeiro, foram extraidos 29 HF, incluindo nove recursos de intensidade, oito recursos geométricos e doze recursos de textura baseados na matriz de coocorrência no nível de cinza (GLCM). Em seguida, foiram treinadas CNNs 3D modificadas a partir de três arquiteturas 2D CNN (AlexNet, VGG-16 Net e Multi-crop Net) para extrair os recursos CNN aprendidos na camada de saída. Para cada CNN 3D, os recursos da CNN combinados com o 29 HF foram usados como entrada para a máquina de vetores de suporte (MVS), juntamente com o método de seleção de recurso sequencial para frente (SFS) para selecionar o subconjunto de recursos ideal e construir os classificadores. A base de pacientes inclui 431 nódulos malignos e 795 nódulos benignos extraídos do banco de dados LIDC/IDRI. Foram obtidos os resultados de 0,9306, 88,58%, 82,60% e 91,82% para AUC, precisão, sensibilidade e especificidade, respectivamente.

Por fim, o trabalho de (ZHANG et al., 2019) apresenta um novo método de classificação para nódulos pulmonares com base em características híbridas de imagens de tomografia computadorizada (TC). O método fundiu recursos de rede profunda de caminhos duplos 3D (DPN), recursos de textura baseados em padrão binário local (LBP) e histograma de recursos de forma baseados em gradientes orientados (HOG) para caracterizar nódulos pulmonares. Foi utilizado o conjunto de dados LUng Nodule Analysis 2016 (LUNA16) disponível ao público com 1004 nódulos, atingindo uma área abaixo da curva de característica operacional do receptor (AUC) de 0,9687 e precisão de 93,78%.

Estes são exemplos de trabalhos que foram desenvolvidos para o diagnóstico de nódulo pulmonar. Vale ressaltar também, que assim como a metodologia proposta, alguns trabalhos utilizam Índices de Diversidade para descrição de textura, como é o caso do trabalho apresenta em (NASCIMENTO, 2012). Nos trabalhos apresentados, alguns pontos como, poucos casos de análise, discrepância entre os valores de sensibilidade e especificidade, são problemas comuns de sistemas dessa natureza. Sendo assim, a metodologia propostas buscará explorar estas deficiências com intuito de melhorá-las.

Este trabalho está organizado da seguinte forma. No Capítulo 2, apresentamos a fundamentação teórica necessária para a melhor compreensão do trabalho, no capítulo 3, apresentamos a metodologia utilizada para classificar os nódulos extraídos da TC como benigno e maligno, utilizando a extração de características por meio de índices taxonômicos, algoritmo genético e classificação por meio da Maquina de Vetores e Suporte(MVS). No capítulo 4 mostramos e discutimos os resultados alcançados através da metodologia proposta. Finalmente, no capítulo 5, apresentamos as considerações finais sobre este trabalho.

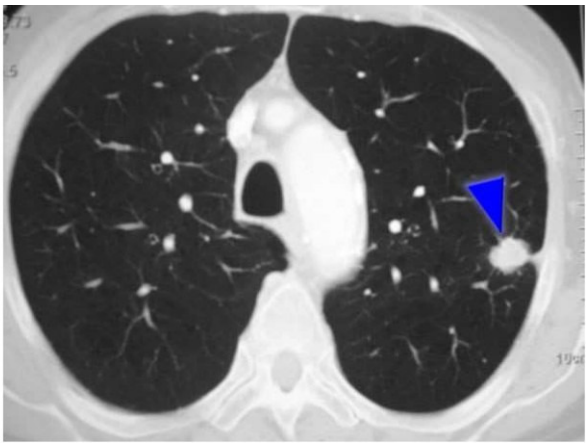
2 Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica utilizada no desenvolvimento deste trabalho e necessária para compreensão das técnicas utilizadas para alcançar os objetivos.

2.1 Nódulo Pulmonar Solitário

O Nódulo Pulmonar Solitário é uma lesão sólida, geralmente arredondada, menor que 3 cm de diâmetro (lesões maiores que 3 cm são denominadas "massas"), cercada de pulmão normal que pode ser de natureza benigna ou maligna (NASCIMENTO, 2012). A Figura 1 mostra um exemplo de um nódulo pulmonar em uma fatia de uma tomografia computadorizada.

Figura 1 – Exemplo de nódulo pulmonar



Fonte: (FISHMAN et al., 2004)

Algumas das características dos nódulos pulmonares que ajudam a inferir sobre a probabilidade de benignidade e malignidade incluem (CHATE; FUNARI, 2011):

1. **Tamanho:** Os dados da literatura demonstram claramente que a probabilidade de malignidade aumenta com o aumento do tamanho do nódulo pulmonar.
2. **Localização:** Os cânceres de pulmão são mais frequentes nos lobos superiores.
3. **Existência ou Não de Calcificação e Gordura:** A existência de calcificação constitui evidência quase certa de benignidade com raríssimas exceções. A identificação de gordura no interior de um nódulo pulmonar também é uma característica fortemente sugestiva de benignidade quase sem exceções.

4. **Tempo de Duplicação:** Um nódulo com tempo de duplicação muito curto (por exemplo, inferior a um mês) ou muito longo (classicamente, superior a 450 dias) tem maior probabilidade de ser benigno. Ao contrário, um nódulo pulmonar cujo tempo de duplicação estiver entre esses limites tem maior chance de revelar-se maligno.

Todavia, o diagnóstico definitivo de malignidade é dado somente pelo exame citopatológico do material obtido por procedimentos que estão se tornando de menor morbidade, como a biopsia transbrônquica e transtorácica. (CHATE; FUNARI, 2011):

2.2 Tomografia Computadorizada

A tomografia computadorizada é um exame simples, capaz de obter imagens em tons de cinza de "fatias" de partes do corpo ou de órgãos selecionados, as quais são geradas graças ao processamento por um computador de uma sucessão de imagens de raios-X de alta resolução em diversos segmentos sucessivos de partes do corpo ou de órgãos. Hoje em dia existem vários modelos de aparelhos de tomografia computadorizada e o funcionamento deles pode diferir um pouco uns de outros, mas todos têm em comum o fato de se utilizarem dos raios-X para obterem imagens do interior do corpo (ABCMED, 2019).

A tomografia computadorizada baseia-se nos mesmos princípios técnicos que a radiografia tradicional, e na verdade é uma evolução técnica dela, que usa uma radiação maior e toma imagens fatiadas dos segmentos que examina, as quais o médico superpõe imaginativamente para obter uma visão tridimensional. Em alguns casos há necessidade de se utilizar um contraste injetável, a fim de aumentar a capacidade diagnóstica. As imagens da tomografia podem ser tomadas em dois planos básicos: o axial (perpendicular ao maior eixo do corpo) e o coronal (paralelo à sutura coronal do crânio) e permitem reconstruções no plano sagital (paralelo à sutura sagital do crânio) e tridimensionais (ABCMED, 2019).

A tomografia computadorizada é usada para detectar tumores, fraturas, obstruções circulatórias, alterações nas estruturas orgânicas e outras anomalias teciduais, sendo mais precisa para tecidos moles que as simples radiografias. Hoje em dia a tomografia computadorizada vem apresentando menor volume de exames em comparado a ressonância magnética em virtude de duas grandes vantagens dessa última: imagens com maior definição e o fato de não usar energia radioativa (ABCMED, 2019).

2.3 Quantização Uniforme

Uma imagem digital é discretizada espacialmente em x e y, e também em amplitude (intensidade luminosa). A discretização em amplitude é conhecida como quantização.(GONZALEZ; WOODS, 1992)

A quantização uniforme consiste em dividir a escala de cinza da imagem em intervalos iguais, em que cada intervalo é mapeado para um valor de cinza na imagem quantizada, de modo que a escala de cinza da imagem quantizada é dada por [0, L'- 1], sendo o nível de cinza da imagem quantizada (L') menor do que da imagem original (L), ou seja, L' < L(GONZALEZ; WOODS, 1992).

A expressão para calcular esse mapeamento é:

$$q(i,j) = (2^b - 1)\frac{p(i,j) - I_{min}}{I_{max} - I_{min}} \tag{2.1}$$

onde q(i,j) é o nível de cinza do pixel (i, j) da nova imagem (quantizada), $p(i, j)$ é o nível de cinza do pixel (i, j) da imagem original, $[I_{max} - I_{min}]$ é a escala de cinza da imagem original, e b é o número de bits necessário para armazenar cada pixel da imagem quantizada.

A técnica de quantização uniforme será aplicada na fase de extração de características para cada nódulo, como um preprocessamento, antes da extração de características em si, para investigar a descrição das informações de textura dos nódulos em imagens com diferentes níveis de cinza.

2.4 Análise de Textura

A textura é definida como a característica de uma região relacionada a coeficientes de uniformidade, densidade, aspereza, regularidade, intensidade, entre outras características da imagem (HARALICK; SHANMUGAM; DINSTEIN, 1973).

A análise de textura é relevante em imagens digitas, uma vez que possibilita distinguir regiões da imagem que apresentam as mesmas características de padrões (CONCI; AZEVEDO; LETA, 2008). Uma forma clássica de quantificação da textura numa imagem em níveis de cinza é a abordagem estatística, a qual propicia a descrição da textura através das regras estatísticas que regem tanto a distribuição quanto à relação entre os diferentes níveis de cinza de uma região da imagem.

Neste trabalho foi proposta a descrição da textura dos tecidos de regiões dos nódulos pulmonares solitários através do Índice de Diversidade Taxonômica (Δ) e da Distância Média Filogenética (MPD), que são medidas estatísticas.

2.4.1 Índice de Diversidade

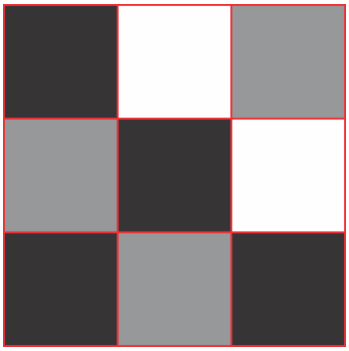
A diversidade é um termo muito utilizado na área da ecologia. O seu objetivo é informar a variedade de espécies presentes em uma comunidade ou área. O conceito de comunidade é descrito como um conjunto de espécies que ocorrem em um determinado lugar e tempo (MAGURRAN, 2013). As medições como a variância e o desvio padrão

que são calculadas em estudos estatísticos, apresentam valores que medem a variabilidade quantitativa, enquanto que os índices de diversidade descrevem a variabilidade qualitativa.

Para medir a diversidade temos duas componentes: a riqueza de espécies, que consiste no número de espécies encontradas em determinada região, e a abundância relativa, que é o número de indivíduos de uma determinada espécie existentes numa dada área (VITT; PIANKA, 2014). O resultado do cálculo, para qualquer índice de diversidade, é representado por um único valor (SANTOS, 2009). As medidas de diversidade de espécies são geralmente úteis para comparar padrões em diferentes áreas.

A forma mais simples da aplicação do índice de diversidade em imagens é quando a comunidade representa uma imagem ou região da mesma, as espécies sendo os níveis de cinza e os indivíduos sendo os pixels (SOUSA et al., 2011). A Figura 2 mostra um exemplo demonstrativo.

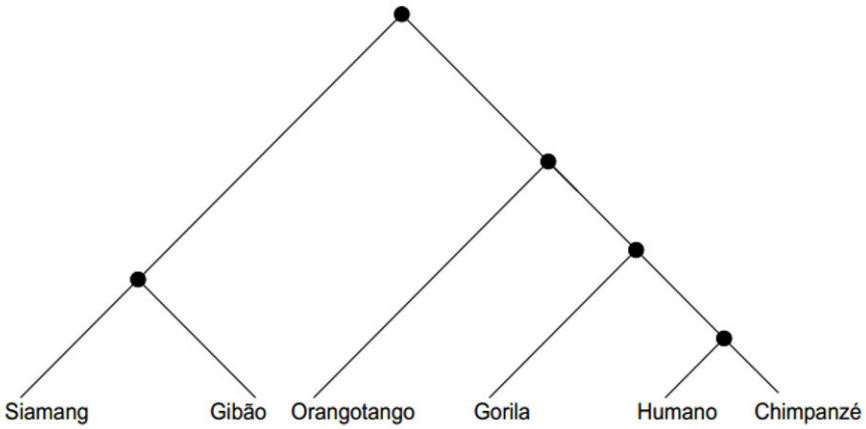
Figura 2 – Exemplo de uma imagem que possui três níveis de cinza (espécies) que é preto, o cinza e o branco. A quantidade de pixels (indivíduos) de preto é 4, de cinza é 3 e de branco é 2.



2.4.2 Diversidade Filogenética

A filogenia é um ramo da biologia responsável pelo estudo das relações evolutivas entre as espécies, pela verificação dos relacionamentos entre elas, a fim de determinar possíveis ancestrais comuns. Uma árvore filogenética, ou simplesmente filogenia, é uma árvore onde as folhas representam os organismos e os nós internos representam supostos ancestrais. As arestas da árvore denotam as relações evolutivas. Na Figura 3 temos um exemplo de árvore filogenética, em que se verifica o relacionamento entre espécies de macacos e a espécie humana, onde podemos ver que o homem e o chimpanzé são geneticamente mais próximos que os outros pares presentes na árvore (ARAÚJO, 2003).

Figura 3 – Representação de uma árvore filogenética para alguns primatas.



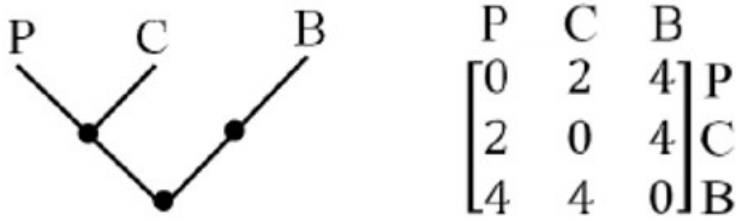
Fonte: (ARAÚJO, 2003)

De maneira geral, a diversidade não pode ser medida apenas com a utilização de dados como a abundância e a riqueza de espécies, cada vez mais o parâmetro filogenético vem sendo inserido neste cálculo (CLARKE; WARWICK, 1998). A diversidade filogenética é uma medida da diversidade de uma comunidade que incorpora as relações filogenéticas das espécies (MAGURRAN, 2013). A combinação da abundância das espécies com a proximidade filogenética para gerar um índice de diversidade é denotada diversidade taxonômica (SILVA; BATALHA, 2006). A taxonomia é a ciência que lida com a classificação (criação de novas taxas), identificação (alocação de linhagens dentro de espécies conhecidas) e nomenclatura (VANDAMME et al., 1996).

Clarke e Warwick (1998) desenvolveram um método para mensurar a diversidade taxonômica muito sensível a perturbações ambientais e apropriado para avaliar as diferenças entre comunidades. Uma comunidade em que as espécies estão distribuídas em muitos gêneros deve apresentar uma diversidade maior que uma comunidade em que a maioria das espécies pertence a um mesmo gênero (MAGURRAN, 2013).

A diversidade taxonômica é baseada no conjunto das distâncias entre pares de espécies acumuladas a partir das árvores taxonômicas (RICOTTA, 2004). A Figura 4 apresenta uma ilustração de uma árvore taxonômica em que as folhas são espécies e a soma da quantidade de arestas que ligam determinado par de espécies é informada pela matriz.

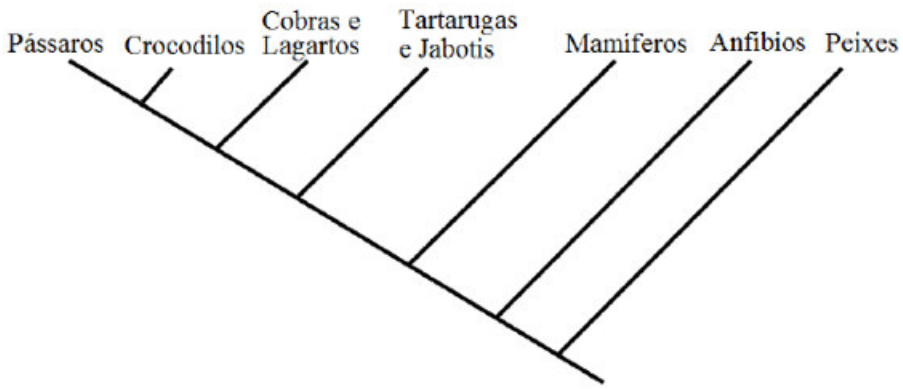
Figura 4 – Exemplo representando uma imagem em uma árvore taxonômica (à esquerda) e sua matriz de distância (à direita).



Fonte: (OLIVEIRA et al., 2013)

Uma das formas de representar a árvore filogenética é através do cladograma, que é um diagrama representativo das relações ancestrais entre organismos. Para este trabalho foi utilizada a topologia de um cladograma mais específico, o enraizado na forma de cladograma inclinado (VIANA; CEARÁ, 2007). Este tipo de árvore é mostrado na Figura 6, que descreve a sequência evolutiva de alguns tetrápodes (vertebrados terrestres possuidores de quatro membros).

Figura 5 – Árvore filogenética enraizada na forma de cladograma inclinado.



Fonte: (VIANA; CEARÁ, 2007)

2.4.3 Índices Taxonômicos

O cálculo entre dois organismos escolhidos aleatoriamente em uma filogenia existente em uma comunidade é apresentado por índices de Diversidade Taxonômica e Distinção Taxonômica (CLARKE; WARWICK, 1998).

O Índice de Diversidade Taxonômica (Δ) considera a abundância das espécies e a relação taxonômica entre elas, assim, o seu valor expressa a distância taxonômica média

entre quaisquer dois indivíduos, escolhidos na amostra ao acaso (GORENSTEIN, 2009).

$$\Delta = \frac{\sum\sum_{i<j} w_{ij}x_i x_j}{[\frac{n(n-1)}{2}]} \tag{2.2}$$

onde $x_i (i = 1, ..., s)$ é a abundância da $i - ésima$ espécie, n é o número total de espécies e $_{ij}$ é a distância da espécie i à espécie j na classificação taxonômica.

2.4.4 Distância Filogenética Média (Mean phylogenetic distance- MPD)

O MPD é a distância filogenética média entre todas as combinações de pares de espécies. Ela nos dá um valor geral da estrutura filogenética da comunidade (WEBB, 2000).

$$MPD = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}p_i p_j}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_i p_j} \tag{2.3}$$

Onde $d_{i,j}$ é a distância filogenética entre i e j, e p_i, p_j são 0/1 para ausência/presença de espécies. Com essa média temos uma visão geral da comunidade em um único valor através da analises das espécies que estão presente/ausente mas sem levar em consideração a quantidade de indivíduos de cada.

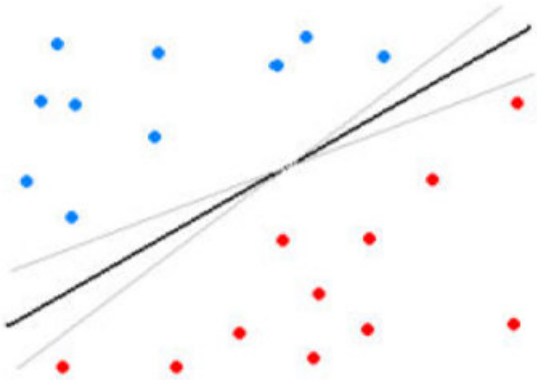
2.5 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (MVS) é uma técnica de aprendizagem, usada para estimar uma função que classifique dados de entrada em duas classes. O princípio básico é a construção de um hiperplano como superfície de decisão, cuja margem de separação entre as classes seja máxima (VAPNIK; VAPNIK, 1998). Por hiperplano entende-se uma superfície de separação de duas regiões em um espaço multidimensional, em que o número de dimensões pode ser, até, infinito.

A Figura 6 mostra em duas dimensões, para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha mais escura), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

Há casos em que podem existir vários possíveis hiperplanos de separação, mas MVS busca apenas encontrar o que maximize a margem entre os exemplos de treinamento (NASCIMENTO, 2012).

Figura 6 – Separação entre duas classes através de hiperplanos.



Fonte: (NASCIMENTO, 2012)

2.6 Algoritmo Genético

Algoritmos Genéticos, AGs, são métodos de otimização e busca inspirados nos mecanismos de evolução de populações de seres vivos. Foram introduzidos por John Holland e popularizados por um dos seus alunos, David Goldberg. Estes algoritmos seguem o princípio da seleção natural e sobrevivência do mais apto, declarado em 1859 pelo naturalista e fisiologista inglês Charles Darwin em seu livro A Origem das Espécies. De acordo com Charles Darwin, " Quanto melhor um indivíduo se adaptar ao seu meio ambiente, maior será sua chance de sobreviver e gerar descendentes". Otimização é a busca da melhor solução para um dado problema. Consiste em tentar várias soluções e utilizar a informação obtida neste processo de forma a encontrar soluções cada vez melhores. Um exemplo simples de otimização é a melhoria da imagem das televisões com antena acoplada no próprio aparelho. Através do ajuste manual da antena, várias soluções são testadas, guiadas pela qualidade de imagem obtida na TV, até a obtenção de uma resposta ótima, ou seja, uma boa imagem (LACERDA; CARVALHO, 1999).

As técnicas de busca e otimização, geralmente, apresentam:

- Um espaço de busca, onde estão todas as possíveis soluções do problema;
- Uma função objetivo (algum as vezes chamada de função de aptidão na literatura de AGs), que é utilizada para avaliar as soluções produzidas, associando a cada uma delas uma nota. Em termos matemáticos, a otimização consiste em achar a solução que corresponda ao ponto de máximo ou mínimo da função objetivo (LACERDA; CARVALHO, 1999).

2.7 Validação dos Resultados

Em um sistema de reconhecimento de padrões relacionado à área médica, os resultados dos testes de classificação em relação ao diagnóstico podem ser divididos em quatro grupos:

- O teste é positivo e o paciente tem a doença – Verdadeiro Positivo (VP);
- O teste é positivo e o paciente não tem a doença – Falso Positivo (FP);
- O teste é negativo e o paciente tem a doença – Falso Negativo (FN);
- O teste é negativo e o paciente não tem a doença – Verdadeiro Negativo (VN).

Para avaliar o desempenho do classificador, é comum utilizar o cálculo de algumas estatísticas como Sensibilidade (SE), Especificidade (ES) e Acurácia (AC) (BLAND, 2000).

A sensibilidade de um teste é definida pela proporção de pessoas com a doença de interesse, cujo resultado é positivo. Indica quão bom é o teste para identificar os indivíduos doentes.

$$SE = \frac{VP}{VP + FN} \tag{2.4}$$

A especificidade de um teste é a proporção de pessoas sem a doença cujo resultado é negativo. Indica quão bom é o teste para identificar os indivíduos não doentes.

$$ES = \frac{VN}{VN + FP} \tag{2.5}$$

A taxa de classificação correta (acurácia) é definida como a razão entre o número de casos na amostra em estudo que foram classificados corretamente e o número total de casos na amostra em estudo.

$$AC = \frac{VP + VN}{VP + FN + VN + FP} \tag{2.6}$$

Com essas metricas é possivel calcular tambem a curva ROC, que é a representação gráfica dos pares sensibilidade ou FVP (ordenadas) e 1- especificidade ou FFP (abcissas), resultantes da variação do valor de corte ao longo de um eixo de decisão, x, a representação gráfica assim resultante é designada por curva ROC no plano unitário. Com efeito, uma curva ROC é uma descrição empírica da capacidade do sistema de diagnóstico poder discriminar entre dois estados num universo, onde cada ponto da curva representa um compromisso diferente entre a FVP e a FFP que pode ser adquirido pela adopção de

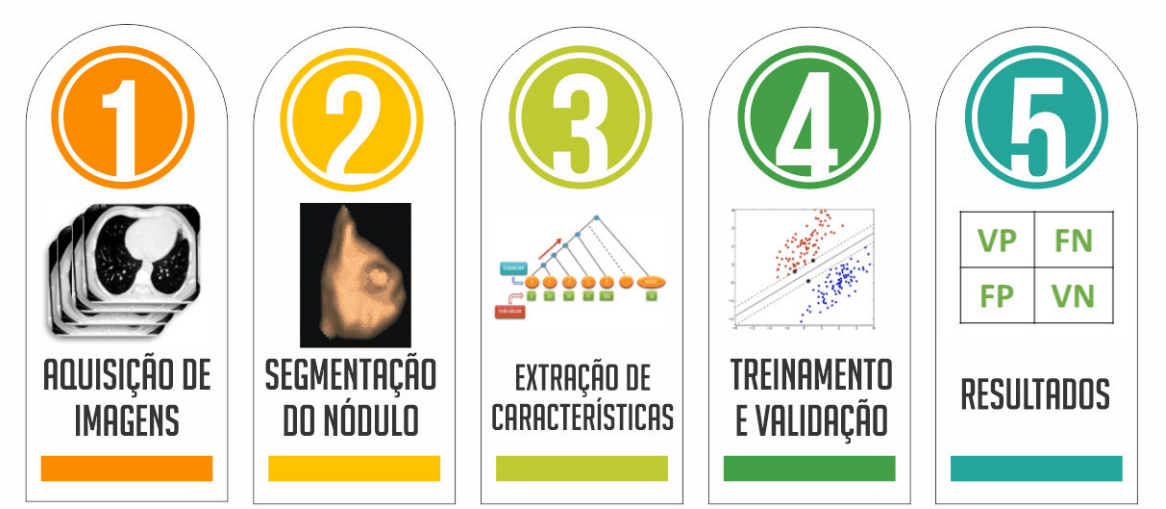
um diferente valor de corte de anormalidade ou nível crítico de confiança no processo de decisão (BRAGA, 2001).

A sensibilidade, a especificidade, a acurácia e a curva ROC foram usadas para avaliar o desempenho da metodologia desenvolvida neste trabalho, considerando nódulos pulmonares malignos corretamente classificados como verdadeiros positivos.

3 Metodologia

Neste capítulo são descritas as etapas utilizadas na metodologia proposta para a classificação de nódulos pulmonares em exames de tomografia computadorizada. A metodologia está dividida em cinco etapas como descrita na Figura 7. Em síntese, a primeira etapa é a aquisição das imagens que foram obtidas da base de dados de imagens de exames de tomografia computadorizada LIDC-IDRI. Na segunda etapa é realizada a segmentação dos nódulos. Na terceira, é feita a extração de características dos nódulos pulmonares, utilizando os índices taxonômicos. Na quarta etapa é realizado treinamento e validação utilizando o MVS e o algorítimo genético (AG). E por fim, na quinta etapa é realizada a validação dos resultados.

Figura 7 – Metodologia Proposta.



3.1 Aquisição de Imagens

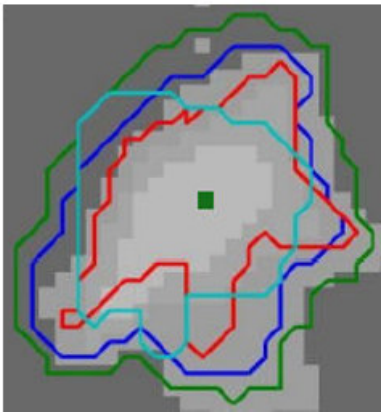
A base de dados utilizada neste trabalho é a LIDC-IDRI (ARMATO; GEOFFRE; AL., 2011), disponibilizada na internet como resultado de uma associação entre Lung Image Database Consortium e a Image Database Resource Initiative com 833 exames de tomografia computadorizada.

Na base LIDC-IDRI, as imagens estão no formato DICOM e possuem 16 bits por voxel. A base fornece um arquivo em formato XML com a informação do contorno ao longo das fatias, além de algumas características como esfericidade, textura, malignidade e etc (a estas é indicado um valor de 1 a 5), para aqueles nódulos pulmonares maiores que 3 mm, e apenas a informação sobre o centróide para aqueles inferiores a 3 mm.

O processo de anotação dos nódulos da base LIDC-IDRI foi feito por quatro especialistas, e em duas fases. Na primeira, cada radiologista analisou os exames de forma independente. Na segunda, os resultados das quatro análises da primeira fase foram apresentados juntos para cada radiologista. Durante essa etapa, eles analisaram e refizeram livremente suas anotações.

Não há imposição para que haja consenso, todos os nódulos indicados pela revisão dos radiologistas são apurados e gravados. Sendo assim, é possível ter diferentes diagnósticos para um mesmo nódulo. Considera-se, então, neste trabalho, apenas uma instância por nódulo, objetivando minimizar o impacto da subjetividade nos exames. No entanto, não existe nenhuma indicação na anotação dos radiologistas (arquivo XML) sobre quais informações se referem ao mesmo nódulo. Para esta tarefa, então, calcula-se o ponto central dos nódulos posteriormente verificando se as coordenadas desse ponto se encontram na região de um nódulo apurado por outro especialista. A Figura 8 ilustra o processo:

Figura 8 – Ilustração do resumo das marcações dos nódulos.



Fonte: (NASCIMENTO, 2012)

Na Figura 8, as linhas coloridas representam os contornos definidos pelos especialistas individualmente. O quadrado verde, no centro, refere-se ao centróide calculado para o contorno da mesma cor. Conforme o cálculo, as coordenadas desse centróide se encontram nas áreas delimitadas por outros especialistas. Dessa forma, considera-se, neste trabalho, que se trata do mesmo nódulo e, portanto, só deve ser aceita uma instância, referente àquela marcação do nódulo que possuir a maior área de contorno. Após o cálculo de quais nódulos foram anotados por mais de um especialista, e de selecionar quais as instâncias correspondentes que serão utilizadas, é feito o resumo do diagnóstico quanto à malignidade ou benignidade. O diagnóstico já está presente para cada nódulo gravado na base, em uma escala de malignidade de cinco níveis representados no arquivo XML por números de 1 a 5 (“altamente improvável”, “moderadamente improvável”, “indeterminado”, “moderadamente provável”, ou “altamente provável”, respectivamente). Para o resumo, então, utilizam-se as informações conforme levantamento da etapa anterior para que seja

efetuado o cálculo segundo apresentado em (SA; DS; JD., 2009), em que os valores das características pertencentes ao mesmo nódulo são reduzidos a um único valor através do cálculo da moda ou mediana.

Em (SA; DS; JD., 2009) é proposto o mesmo cálculo para resumir todas as características, mas para o presente trabalho somente a característica de malignidade é importante. Portanto, é a única considerada e computada. Nódulos com taxa de malignidade de 1 ou 2 foram considerados benignos, com taxa de 4 ou 5 foram considerados malignos, os nódulos com taxa 3 foram considerados indeterminados. O método se baseia no cálculo da moda e só no caso de inexistência de moda, ou decorrência de bimodalidade é que utiliza-se o cálculo da mediana. Como se tratam de números inteiros, e pode ocorrer um resultado fracionado, para a mediana, deve-se sempre arredondar o resultado para baixo. Ao total, após as etapas do resumo feito para os 833 exames presentes na base LIDC-IDRI, foram obtidos 2.393 nódulos, sendo 1.011 benignos, 394 malignos e 988 indeterminados. Nesse trabalho utilizaremos apenas o nódulos malignos e benignos, ficando assim com uma base de 1.405 nódulos.

3.2 Segmentação dos Nódulos

Para a segmentação dos nódulos, são obtidas informações do seu contorno de um arquivo XML que contém as coordenadas dos nódulos segundo critério de análise de cada especialista. No entanto, segmentação utilizada nesse trabalho segue o resumo apresentado na Seção 3.1, em que somente a maior delimitação é escolhida para representar a instância dos nódulos descritos por até quatro especialistas.

3.3 Extração de Características

Após a segmentação dos nódulos, eles são submetidos a etapa de extração de características. Essa fase é dividida em subetapas:

1. Quantizar o nódulo: Os nódulos são quantizados em dois níveis de cinza: 12 e 8 bits. Desta forma, nós usamos os nódulos com 16 (imagem original), 12 e 8 bits. A vantagem desta estratégia é poder analisar o nódulo em diferentes variações de níveis de cinza.

2. Analisar regiões do nódulo: Com o intuito de analisar mais detalhadamente os nódulos pulmonares foram utilizadas duas abordagens: máscara interna e externa. Com elas é possível analisar as áreas próximas as bordas e as áreas mais internas dos nódulos separadamente. Nós usamos a mesma ideia proposta por (OLIVEIRA et al., 2015).

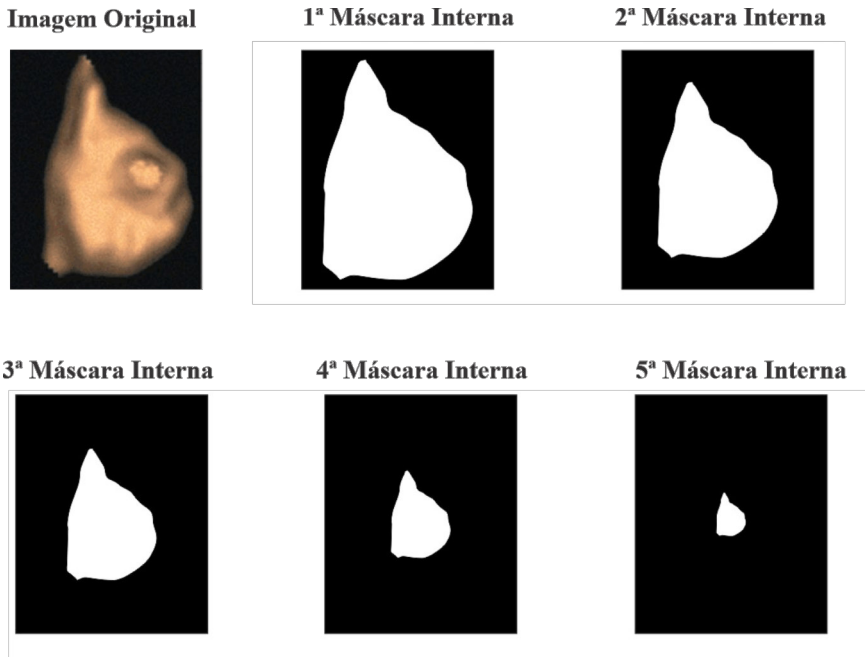
3. Árvores filogenéticas: Nós usamos três arquiteturas de árvores filogenéticas com o objetivo de modificar as relações entre espécies, o que influencia diretamente no cálculo dos índices utilizados nesse trabalho.

4. Índices de diversidade: Para analisar a textura dos nódulos foram utilizados dois índices de diversidade: índice de diversidade taxonômico e distância média filogenética. Estes índices são baseados em distância filogenética, que afere o grau de parentesco entre as espécies.

3.3.1 Abordagem de máscara interna e externa

Essa abordagem visa encontrar padrões de diversidade nas áreas próximas à fronteira das regiões e nas áreas internas (OLIVEIRA et al., 2015). Essas regiões foram geradas através de máscaras que são imagens binárias. A primeira máscara interna foi criada com a binarização do volume de interesse quantizado (VOI), a segunda máscara interna é baseada em sucessivas reduções da escala do VOI em relação à primeira, mantendo o centro de massa. As máscaras sucessoras foram adquiridas dos seus seguidores anteriores para as mais internas. Definimos um valor de 20% para a diminuição da escala, pois foi verificado em testes que os melhores resultados foram alcançados usando cinco máscaras de imagem com essa proporção de escala. O esquema do procedimento para geração das máscaras e, consequentemente, de suas áreas de interesse é apresentado na Figura 9.

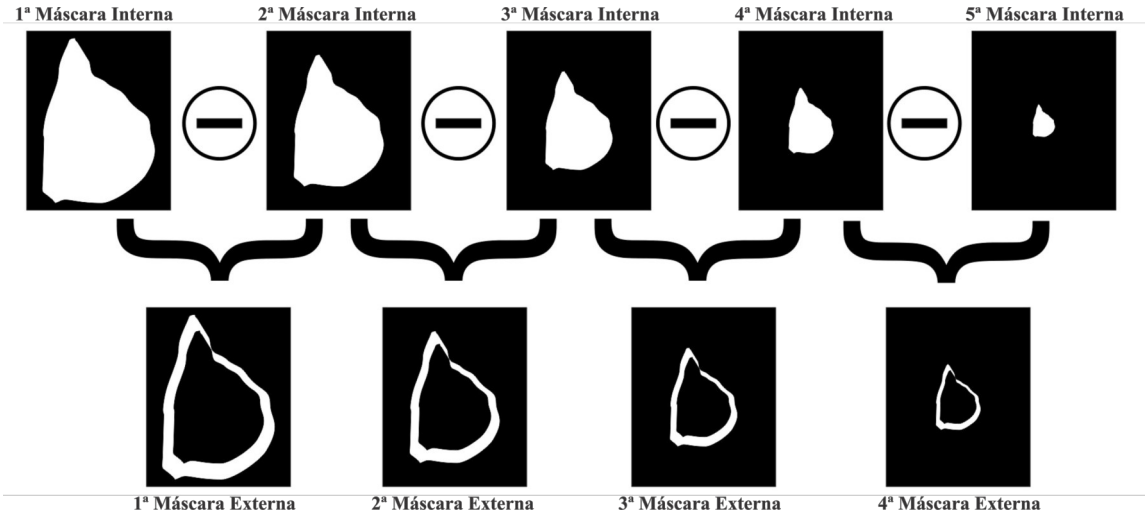
Figura 9 – Abordagem de máscara interna.



As máscaras externas são determinadas pela diferença entre as máscaras internas, onde a primeira máscara externa é determinada pela diferença entre a primeira e a segunda

máscara interna, e assim por diante.Na Figura 10, são demonstrados os passos para criação das máscaras externas.

Figura 10 – Abordagem de máscara externa.



Ao final dessa subetapa se tem um total de 9 máscaras (5 internas e 4 externas) para cada nódulo pulmonar. Assim pode ser feita uma análise de diversidade mais detalhada em cada nódulo.

3.3.2 Árvores filogenéticas

A árvore filogenética agrupada com o Δ e o MPD são usados na Biologia para comparar padrões de comportamento das espécies em diferentes áreas. A fim de implementar esta idéia, o primeiro passo é fazer uma correspondência entre os termos utilizados na biologia e os utilizados em nossa metodologia. A Tabela 1 apresenta esta correspondência.

Tabela 1 – Correspondência entre termos da Biologia e nosso trabalho

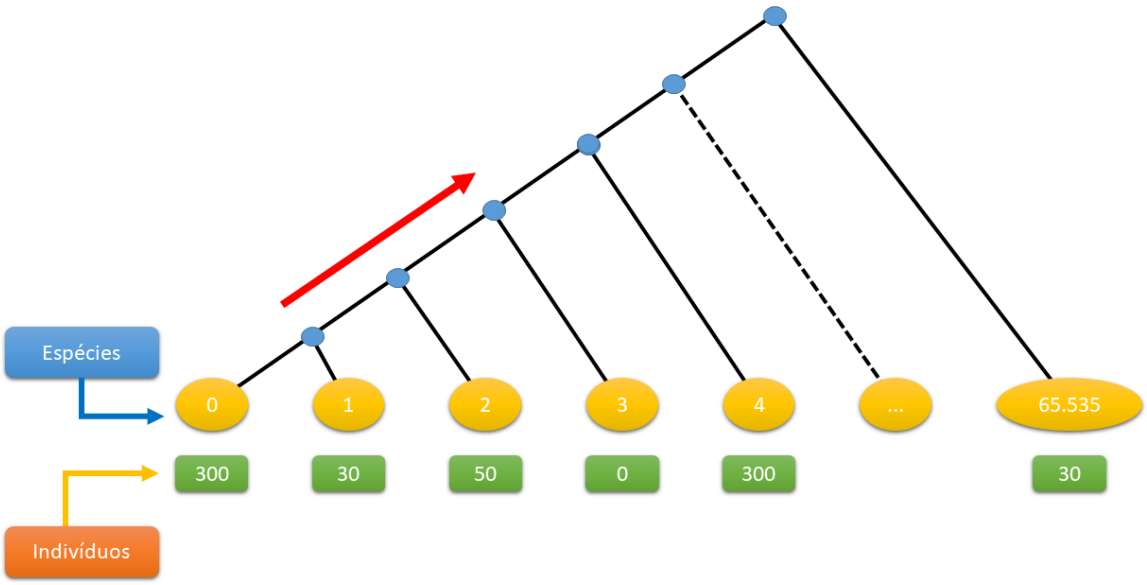
Biologia	Metodologia proposta
Comunidade	Região de interesse da imagem de TC
Espécies	Número máximo de unidade de Hounsfield (UH) na região
Riqueza de espécies: número de espécies encontradas em uma determinada região	Riqueza de espécies: número de voxels encontrados na região
Indivíduos	Número de voxels de uma espécie particular encontrados na região
Abundância Relativa: número de indivíduos de certas espécies existentes em uma determinada área	Número de voxels encontrados na região, que têm o mesmo valor de UH (espécie)

3.3.2.1 Árvore enraizada na forma de cladograma inclinado

Após a geração das abordagens de máscara interna e externa (Seção 3.3.1), são geradas as árvores para cada uma das áreas. Os índices utilizados neste trabalho (MPD e Δ) são baseados nas árvores filogenéticas que possuem três fatores essenciais para aplicação: número de espécies, número de indivíduos e a estrutura de ligação das espécies (quantidade de arestas). Para representar as imagens, foi utilizado o modelo de árvore enraizada na forma de cladograma inclinado.

A Figura 11 ilustra uma árvore filogenética (denominada árvore 1), onde as espécies são os valores de UH, que podem variar entre +32768 e -32768. Nós fizemos uma mudança de escala no valores de UH somente para ficar positivo e tornar mais simples o cálculo dos índices. Esta mudança ocorre quando se desloca o menor valor negativo para que se inicie em zero, podendo variar até 65536 espécies.

Figura 11 – Árvore 1: árvore enraizada na forma de cladograma inclinado.



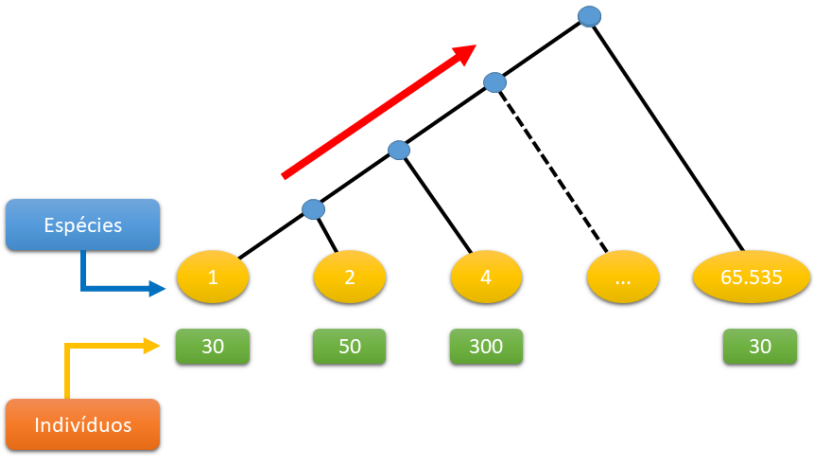
A relação entre as espécies da árvore 1 (Figura 11) é feita no sentido da esquerda para direita como indica a seta vermelha. Assim, a primeira relação é entre a espécie 0 e 1 que possui duas arestas ligando as mesmas, como mostra a Figura 12(a). Na segunda relação são três arestas que ligam as espécies 0 e 2 (Figura 12(b)). Em seguida é a combinação entre as espécies 0 e 3, com quatro arestas (Figura 12(c)). Na última relação é encontrado 65536 arestas entre 0 e 65535 (Figura 12(d)).

Figura 12 – Relações entre as especies(UH).

3.3.2.2 Árvore enraizada na forma de cladograma inclinado excluindo as espécies sem indivíduos

Seguindo a mesma lógica do cálculo dos índices com base na árvore anterior, foi desenvolvida outra arquitetura de árvore que tem como destaque a eliminação das espécies que não possuem indivíduos, resultando consequentemente na reorganização das arestas para as espécies restantes. Supondo que a árvore da Figura 11 tenha somente indivíduos nas espécies 1, 2 e 4, o novo modelo de árvore possui somente essas espécies e segue a mesma arquitetura, como descreve a seguir na Figura 13.

Figura 13 – Árvore 2: modelo criado a partir da árvore 1, excluindo as espécies que não possuem indivíduos.



3.3.2.3 Árvore enraizada na forma de cladograma inclinado modificando as arestas

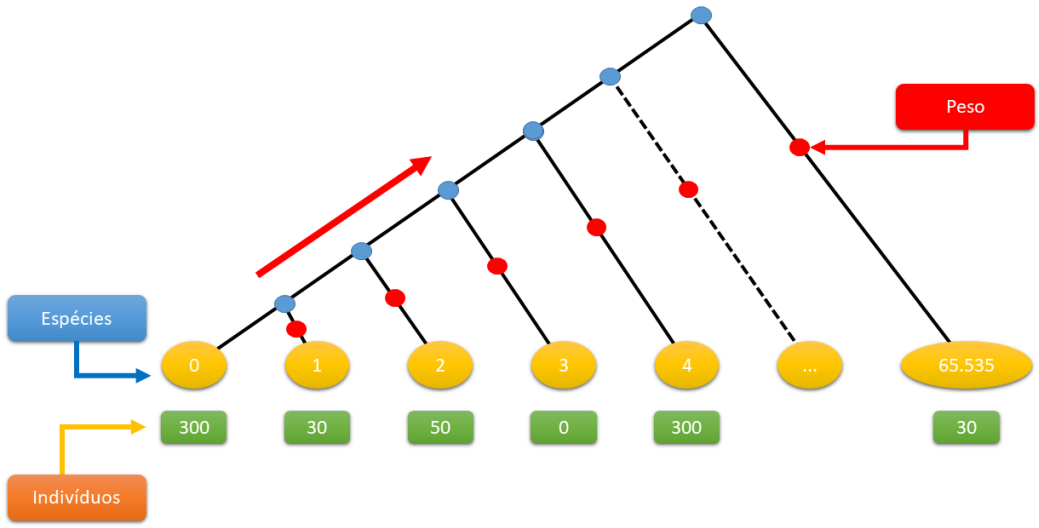
A terceira árvore proposta tem o mesmo processo de combinação entre as espécies da árvore 1, onde a única diferença é na quantidade de arestas do lado direito do nó ancestral entre as espécies. Assim, a Figura 16 descreve o mesmo procedimento da Figura 11, sendo destacado em vermelho as arestas que tinha somente uma ligação na árvore 1 e a primeira combinação de uma espécie com as outras que não tem mudança.

3.4 Treinamento e Validação

A etapa anterior produz um conjunto de vetores de características extraídos de nódulos benignos e malignos. Esses vetores de características serão usados para determinar um modelo de treinamento otimizado pelo algoritmo genético (AG) e pela máquina de vetores de suporte (SVM) (DUDA; HART, 1973). Neste trabalho foi utilizado o AG proposto por (FILHO et al., 2014).

O conjunto de vetores de características gerados na etapa de extração é divido em duas partes: treinamento e teste, com 80% e 20% respectivamente. Como o conjunto de treinamento é desbalanceado (2,5 benignos para cada 1 maligno aproximadamente) o mesmo é submetido ao AG para que seja gerado um modelo de treinamento balanceado para a MVS. O AG criado para otimizar e determinar o melhor modelo possui as seguintes características:

Figura 14 – Árvore 3: modelo criado a partir da árvore 1, modificando as arestas



- Cada geração possui 10 cromossomos
- A seleção é feita através da técnica da roleta
- A elite é formada por um cromossomo
- O cruzamento entre dois cromossomos A e B troca as características que eles não compartilham. Um vetor só pode ser trocado por outro vetor do mesmo tipo de subamostra (treinamento com treinamento, validação com validação). Cada cromossomo muda aleatoriamente entre 1 e 5 de seus vetores.
- A mutação de um cromossomo A substitui aleatoriamente entre 1 e 5 de seus vetores por novos que não pertencem a A.
- Nos casos de convergência precoce, a população é reiniciada, mantendo apenas a elite inalterada.
- A adequação de um cromossomo é calculada através da média aritmética de sensibilidade, especificidade e precisão encontrada na classificação da subamostra de validação. Quanto menor a média, mais apto será o indivíduo e, consequentemente, melhor será o diagnóstico correto do nódulo benigno ou maligno.
- O critério de parada é atingido quando a aptidão do indivíduo mais apto é repetida por 100 gerações consecutivas.

O modelo de treinamento gerado será utilizado no diagnóstico do nódulo benigno ou maligno no conjunto de treinamento.

3.5 Validação dos Resultados

Após a finalização da etapa de treinamento e validação, é necessário validar os resultados e discutir prováveis melhorias. Essa metodologia usa métricas comumente empregadas em sistemas CAD/CADx, e aceitas pela sociedade para análise de desempenho de sistemas baseados em processamento de imagens. Estas métricas são sensibilidade, especificidade e acurácia. Tais métricas têm o objetivo de medir o desempenho da metodologia como satisfatória ou não, além de ajudar a identificar pontos positivos e negativos para melhoria futura deste trabalho na fase de treinamento e teste.

4 Resultados

Neste capitulo serão apresentados os resultados obtidos pela metodologia proposta na classificação de nódulos pulmonares em benigno e maligno.

4.1 Aquisição de Imagens

Foram utilizados 833 exames da base LIDC-IDRI, dos quais foram extraidos 1405 nódulos (1011 benignos e 394 malignos).

4.2 Extração de Características

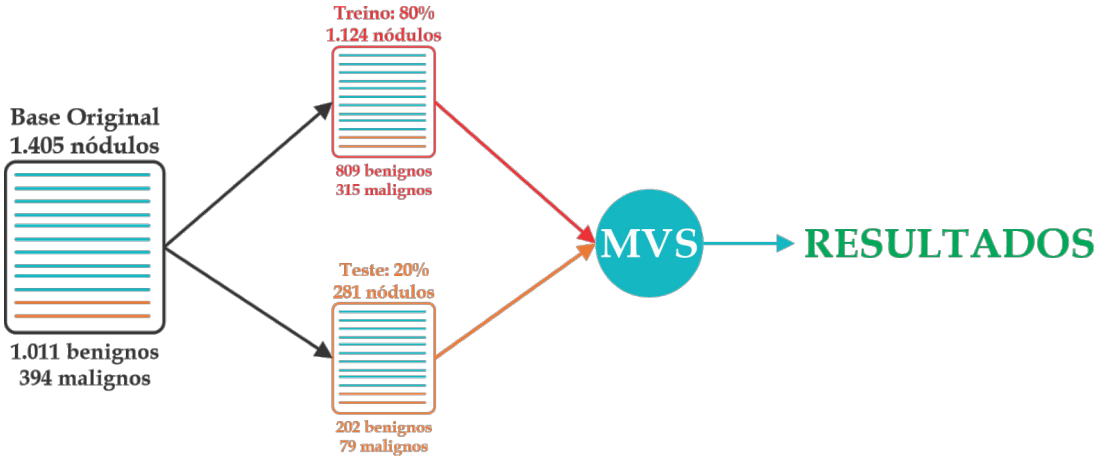
Como dito na Seção 3.3, nós usaremos 3 representações de níveis de cinza para cada nódulo, 16 (imagem original), 12 e 8 bits. Além disso, nós dividimos o nódulo em 9 áreas baseado no conceito de máscara interna e externa. Assim, nós usaremos 27 regiões (3 representações x 9 áreas) para calcular os dois índices de diversidades, Δ e MPD. O número de características a ser usada pelo AG e MVS foram 54 (27 regiões x 2 índices) por nódulo.

4.3 Treino e Validação

Após a etapa de extração de característica a base foi divida aleatoriamente em 80% (1124) para treino e validação, e 20% (281) para teste. A parti disso foram feitos dois testes:

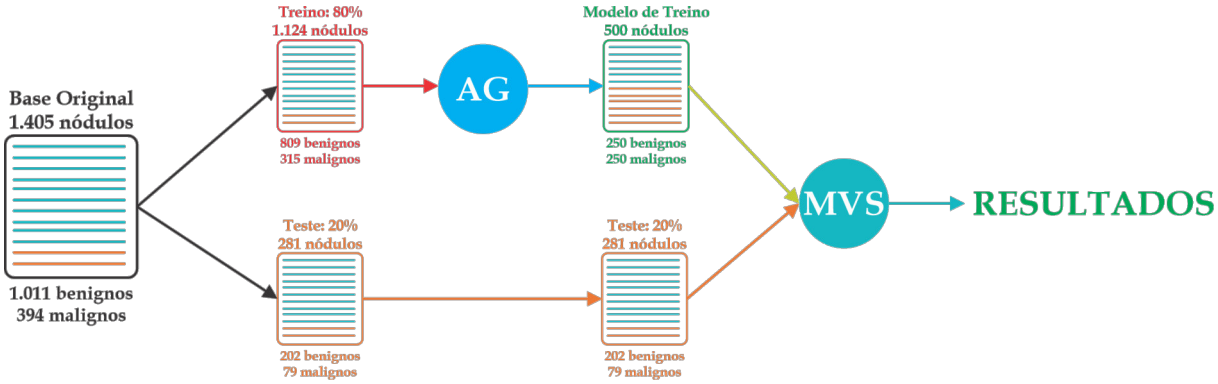
No primeiro teste o conjunto de treino (80%) foi submetido diretamente ao MSV para geração de modelo de treino desbalanceado e em seguida os 20% foram usados para testar o eficacia do modelo.

Figura 15 – Teste 1: Modelo de treino desbalanceado



No segundo teste os 80% da base de treino foram submetidos ao AG para seleção dos melhores indivíduos (nódulos), para que o mesmo gere um modelo balanceado com 500 indivíduos, sendo 250 nódulos malignos e 250 nódulos benignos. Os 20% foram usados para testar o modelo final gerado pelo AG.

Figura 16 – Teste 2: Modelo de treino balanceado



O modelo de treino é balanceado em nossa metodologia porque a utilização de MVS e uma base de treinamento desbalanceada podem resultar em uma classificação tendenciosa em direção às classes com o maior número de indivíduos (DU; CHEN, 2005).

4.4 Classificação

No primeiro teste o conjunto de treinamento (1.124) foi submetido diretamente ao MSV para geração de modelo de treino desbalanceado e em seguida os 20% foram usados para testar o eficacia do modelo. A Tabela 2 mostra os resultados obtidos.

Tabela 2 – Resultado obtido pelo MVS sem balanceamento

Árvore	Acurácia	Sensibilidade	Especificidade
1	82,99%	82,16%	83,47%
2	88,61%	85,67%	89,75%
3	83,84%	83,66%	83,89%

No segundo teste o conjunto de treinamento (1.124) foi submetido ao AG para a seleção do melhor modelo de treinamento. Isso foi realizado gerando um modelo balanceado com 500 nódulos (250 nódulos malignos e 250 nódulos benignos) para treinamento e o restante dos indivíduos (65 nódulos malignos e 559 nódulos benignos) para validar o modelo selecionado. A Tabela 3 mostra os resultados obtidos na seleção do melhor modelo de treinamento para cada árvore.

Tabela 3 – Melhor resultado obtido pelo AG e MVS no conjunto de validação

Árvore	Acurácia	Sensibilidade	Especificidade	Fitness
1	95.51%	100%	94.91%	490.43
2	94.74%	100%	94,06%	488.77
3	86.53%	95.77%	85.35%	459.21

A Tabela 4 apresenta os resultados obtidos usando o modelo selecionado pelo AG em conjunto com o MVS aplicado ao conjunto de testes (79 nódulos malignos e 202 nódulos benignos).

Tabela 4 – Resultados obtidos com nossa metodologia

Árvore	Acurácia	Sensibilidade	Especificidade	ROC
1	91.10%	90.14%	91.20%	0.93
2	**91.81%**	**93.42%**	**91.21%**	**0.94**
3	90.39%	79.76%	94.92%	0.92

Nos experimentos descritos na Tabela 4 foi constatado que a árvore 2 apresenta valores mais altos em todas as métricas, exceto na especificidade, mas com valor acima de 91%. O bom resultado da árvore 2 deve-se ao fato que de acordo com (MELO, 2008) ter conhecimento a priori da diversidade de espécie numa área estudada, é fundamental para a compreensão da natureza. Como nesta árvore estão presentes somente as espécies que tem indivíduos, são excluídos dos cálculos dos índices de diversidade os relacionamentos filogenéticos com as espécies ausentes, ou seja, são relacionadas somente as espécies presentes na região de interesse analisada, diferente das arvores 1 e 3, onde o relacionamento é feito entre todas as especies existentes, estando presente ou ausente na região de interesse analisada, sendo que a arvore 3 é adicionado pesos que torna essas especies ainda mais distintas filogeneticamente.

4.4.1 Comparação com trabalhos relacionados

A comparação com outros trabalhos na área é uma tarefa difícil, uma vez que nenhum dos trabalhos citados neste trabalho forneceu os exames utilizados no treino ou teste. Assim, não foi possível realizar uma avaliação rigorosa do nosso método em relação a outras trabalhos. Nosso objetivo com a Tabela 5 é apenas fornecer uma visão geral dos resultados encontrados nos trabalhos citados em relação ao nosso trabalho.

Tabela 5 – Comparação dos resultados do nosso trabalho com os trabalhos relacionados.

Trabalho	Base de imagem	Amostra Nódulos	Acc %	Sens %	Spec %	ROC
(KUMAR; RAMESH; AL., 2011)	LIDIC-IDRI	—	80.36	76.47	82.05	—
(ZINOVEV et al., 2011)	LIDIC	—	74.1	—	—	0.69
(FARAG et al., 2011)	ELCAP/LI	—	—	86	97	—
(NASCIMENTO, 2012)	LIDC	73	92.78	85.64	97.89	—
(ELIZABETH et al., 2012)	Privada	—	94.44	—	—	—
(AL-ABSI et al., 2012)	JSRT	246	96.00	—	—	—
(OROZCO et al., 2012)	NBIA/ELCAP	—	82.66	96.15	52.16	—
(KREWER; GEIGER; AL., 2013)	LIDC	33	90.91	—	—	—
(OROZCO; VILLEGAS; AL., 2013)	ELCAP e NBIA	128	84	—	—	—
(FILHO et al., 2013)	LIDIC-IDRI	—	97.55	85.91	97.70	—
(PARVEEN; KAVITHA, 2014)	Privada	11	—	91.38	89.56	—
(DANDIL et al., 2014)	Privada	128	90.63	92.3	89.47	—
(AKRAM et al., 2015)	LIDC	—	96.31	—	—	—
(FARAG et al., 2017)	LIDC	1.191	97.5	—	—	—
(NIBALI; HE; WOLLERSHEIM, 2017)	LIDC-IDRI	831	89.35	91.07	88.64	0,94
(MASOOD et al., 2018)	LIDC-IDRI	—	86.02	83.91	89.32	—
(ZHAO et al., 2018)	LIDC	1.018	85.64	—	—	—
(LI et al., 2019)	LIDC-IDRI	—	88.58	82.60	91,82	—
(ZHANG et al., 2019)	LUNA16	1.004	96.87	—	—	—
Nosso trabalho	**LIDC-IDRI**	**1405**	**91.81**	**93.42**	**91.21**	**0.94**

A Tabela 5 mostra uma comparação entre os resultados encontrados neste estudo e alguns dos trabalhos relacionados. É importante enfatizar que, para realizar uma comparação confiável com esses trabalhos anteriores, seria necessário usar o mesmo banco de dados de imagens, os mesmos exames de treinamento e teste e as mesmas configurações para os classificadores, entre outros parâmetros.

Comparando os melhores resultados alcançados em nosso estudo (Árvore 2) com os apresentados na Tabela 5, é possível ver que nossos resultados são promissores. Atingimos resultados acima de 91% para dois tipos de situação: (1) diagnóstico usando apenas recursos de textura e (2) um banco de dados grande e complexo.

Os trabalhos de (NASCIMENTO, 2012; ELIZABETH et al., 2012; AL-ABSI et al., 2012; FILHO et al., 2013) apresentam precisão melhor do que nossa pesquisa; no entanto, nossa pesquisa apresenta uma sensibilidade melhor. Em termos de sistema CADx, a sensibilidade é a métrica mais importante, pois mostra o desempenho do modelo para classificar nódulos malignos corretamente, permitindo a rápida intervenção médica.

As metodologias baseadas em aprendizado profundo (KUMAR; RAMESH; AL., 2011; FARAG et al., 2017; NIBALI; HE; WOLLERSHEIM, 2017; MASOOD et al., 2018;

ZHAO et al., 2018; LI et al., 2019), usam uma quantidade de amostras superior ao nosso, uma vez que os nódulos foram analisados por fatia (2D). No entanto somente os trabalhos de (FARAG et al., 2017; LI et al., 2019) apresentaram uma acurácia melhor que a nossa. Porém devemos salientar que nenhum dos dois utilizaram a o conjunto de dados LIDC-IDRI e os conjuntos de dados utilizados por eles contém menos nódulos.

Tabela 6 – Comparação dos resultados do nosso trabalho com os trabalhos relacionados.

Trabalho	Base de imagem	Amostra Nódulos	Acc %	Sens %	Spec %	ROC
(KUMAR; RAMESH; AL., 2011)	LIDIC-IDRI	—	80.36	76.47	82.05	—
(FILHO et al., 2013)	LIDIC-IDRI	—	97.55	85.91	97.70	—
(NIBALI; HE; WOLLERSHEIM, 2017)	LIDC-IDRI	831	89.35	91.07	88.64	0,94
(MASOOD et al., 2018)	LIDC-IDRI	—	86.02	83.91	89.32	—
(LI et al., 2019)	LIDC-IDRI	—	88.58	82.60	91,82	—
Nosso trabalho	**LIDC-IDRI**	**1405**	**91.81**	**93.42**	**91.21**	**0.94**

Observando a Tabela 6 que contém somente os trabalhos que utilizaram o LIDC-IDRIC como conjunto de dados (KUMAR; RAMESH; AL., 2011; FILHO et al., 2013; NIBALI; HE; WOLLERSHEIM, 2017; MASOOD et al., 2018; LI et al., 2019), somente o trabalho de (FILHO et al., 2013) apresentou uma acurácia maior que a nossa, porem com uma sensibilidade inferior (abaixo dos 90%). E quando levamos em consideração somente a sensibilidade nossa metodologia apresentou resutaldo superior aos demais.

Já os demais trabalhos mencionados na Tabela 5 apresentam resultados mais baixos e menor conjunto de dados de imagens que o nosso. A conclusão obtida indicou o seguinte:

1. O uso de índices de diversidade combinados com árvores filogenéticas foi promissor para a caracterização de texturas de nódulos pulmonares. Estudos anteriores (ROCHA et al., 2014; NUNES; SILVA; PAIVA, 2009; FILHO et al., 2014; OLIVEIRA et al., 2015; CARVALHO; PAIVA; SILVA, 2012; FILHO et al., 2016) indicaram que o índice de diversidade é um bom descritor de textura. Este estudo adicionou conceitos de diversidade de espécies usando árvores filogenéticas como representações e a distância filogenética média (MPD).

2. O uso de quantização uniforme para representar a imagem em diferentes níveis de escala de cinza (8 e 12 bits, além da imagem original) produziu melhores resultados do que usar apenas a imagem original (16 bits).

3. Estudos anteriores sobre o diagnóstico de nódulos pulmonares usaram a análise de forma e textura (HARDIE et al., 2008; JING; BIN; LIANFANG, 2010) para obter bons resultados. Este estudo obteve resultados promissores analisando apenas os recursos de textura. Assim, a adição de mais recursos de forma à metodologia proposta trará melhores resultados.

4. Além disso, este estudo ilustrou que os conceitos propostos por (OLIVEIRA et al., 2015), em que uma análise baseada em região poderia trazer mais informações ao processo de classificação e fornecer mais subsídios ao classificador para obter a decisão correta.

5. O uso do AG em conjunto com o MVS para selecionar um conjunto de indivíduos de uma amostra para gerar o melhor modelo de treinamento foi válido de acordo com os resultados de estudos anteriores de (FILHO et al., 2014) e (SAMPAIO et al., 2015). Além disso, o peso atribuído à sensibilidade métrica obteve modelos com melhor desempenho para classificar nódulos malignos corretamente.

6. Finalmente, é importante destacar que o banco de dados LIDC-IDRI é extremamente complexo e diversificado, contendo inúmeros casos diferentes de nódulos pulmonares. Esse banco de dados possui exames que foram extraídos por vários métodos tomográficos, dificultando a classificação através de sistemas CADx.

5 Conclusões

Avaliou-se a metodologia proposta aplicado a um conjunto de 1.405 nódulos pulmonares da base LIDC-IDRI. Com os resultados obtidos nesse trabalho pode-se concluir que:

1. O alto número de mortes causadas por câncer de pulmão evidencia a importância do desenvolvimento de pesquisas voltadas para o diagnóstico precoce. Isso poderia levar a um tratamento mais apropriado para os pacientes, aumentando assim suas chances de sobrevivência. Com base nisso, ferramentas computacionais que fornecem uma opinião auxiliar ao médicos especialistas também podem ser benéficas para os pacientes.

2. Este estudo apresentou uma metodologia para o diagnóstico de nódulos pulmonares utilizando a distância filogenética média e o índice de diversidade taxonômica. Essas características foram usadas em conjunto com diferentes arquiteturas de árvores filogenéticas e SVM para a classificação do nódulo pulmonar como benigno ou maligno. Por isso, é uma ferramenta útil para médicos especialistas.

3. Os resultados obtidos indicaram o desempenho promissor das técnicas de extração de textura propostas. A criação de uma árvore filogenética foi outro fator importante que levou a bons resultados. O uso dessa árvore contribuiu consideravelmente na discriminação entre nódulos benignos e malignos. Embora o banco de dados de imagens utilizado neste estudo seja robusto e tenha garantido uma grande diversidade de nódulos analisados, são necessários testes adicionais com outros bancos de dados para aprimorar a metodologia proposta e torná-la mais robusta e genérica.

5.1 Trabalhos Futuros

A partir da metodologia apresentada nesse trabalho, percebe-se que utilização de índices de diversidade como descritores de textura se mostrou mais que promissor. Neste trabalho foram utilizados apenas dois índices de diversidade(Δ e MDP), que estão em meio a vários outros, nos dando um leque de combinações possíveis para trabalhos futuros.

Pode-se observar também que a quantização foi utilizada a neste trabalho como pré-processamento para mudar as variações de tonalidades de voxel (espécies), em trabalho futuros pode-se estudar a possibilidade de utilizar outros pré-processamentos.

Nota-se também que a base LIDC-IDRI utilizada, apesar de ter um diversidade e grau de dificuldade altos, é considerada pequena em relação a outras base apresentadas

nos trabalhos relacionados. Pode-se aplicar a metodologia proposta em uma base diferente e comprovar a eficácia da metodologia.

Por fim, pode-se estudar o uso de outros classificadores, além da MVS, com ou sem a ajuda do algoritmo genético para balancear o conjunto de treinamento.

Referências

ABCMED. *Tomografia computadorizada. Como é o exame?* 2019. Acessado em https://www.abc.med.br/p/exames-e-procedimentos/344744/tomografia-computadorizada-como-e-o-exame.htm. Citado na página 20.

AKRAM, S.; JAVED, M. Y.; HUSSAIN, A.; RIAZ, F.; AKRAM, M. U. Intensity-based statistical features for classification of lungs ct scan nodules using artificial intelligence techniques. *Journal of Experimental & Theoretical Artificial Intelligence*, v. 27, n. 6, p. 737–751, 2015. Disponível em: <http://dx.doi.org/10.1080/0952813X.2015.1020526>. Citado 2 vezes nas páginas 16 e 42.

AL-ABSI, H. R.; SAMIR, B. B.; SHABAN, K. B.; SULAIMAN, S. Computer aided diagnosis system based on machine learning techniques for lung cancer. In: IEEE. *Computer & Information Science (ICCIS), 2012 International Conference on*. [S.l.], 2012. v. 1, p. 295–300. Citado 2 vezes nas páginas 15 e 42.

ARAÚJO, G. S. de. *Filogenia de Proteomas*. Tese (Doutorado) — Universidade Federal de Mato Grosso do Sul, 2003. Citado 2 vezes nas páginas 22 e 23.

ARMATO, S. G.; GEOFFRE, M.; AL. et. 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/2145272.html>. Citado na página 29.

BLAND, M. Clinical measurement. *An introduction to medical statistics*, Oxford University Press Oxford, England, v. 3, p. 269–294, 2000. Citado na página 27.

BRAGA, A. *Curvas ROC: aspectos funcionais e aplicações*. Tese (Doutorado), 2001. Citado na página 28.

CARVALHO, P. M. de S.; PAIVA, A. C. de; SILVA, A. C. Classification of breast tissues in mammographic images in mass and non-mass using mcintosh's diversity index and SVM. In: *Machine Learning and Data Mining in Pattern Recognition - 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings*. [s.n.], 2012. p. 482–494. Disponível em: <http://dx.doi.org/10.1007/978-3-642-31537-4_38>. Citado na página 43.

CHATE, R. C.; FUNARI, M. B. d. G. Nódulo pulmonar. *Rev Bras Med*, v. 68, n. 1/2, 2011. Citado 2 vezes nas páginas 19 e 20.

CLARKE, K. R.; WARWICK, R. M. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, v. 35, p. 523–531, 1998. Citado 2 vezes nas páginas 23 e 24.

CONCI, A.; AZEVEDO, E.; LETA, F. *Computaçao Gráfica—Teoria e Prátic, vol. 2*. [S.l.]: Elsevier, New York, NY, USA, 2008. Citado na página 21.

DANDIL, E.; CAKIROGLU, M.; EKSI, Z.; AL. et. Artificial neural network-based classification system for lung nodules on computed tomography scans. *International Conference of Soft Computing and Pattern Recognition*, p. 382–386, 2014. Citado 2 vezes nas páginas 16 e 42.

DU, S.-X.; CHEN, S.-T. Weighted support vector machine for classification. In: IEEE. *Systems, Man and Cybernetics, 2005 IEEE International Conference on*. [S.l.], 2005. v. 4, p. 3866–3871. Citado na página 40.

DUDA, R. O.; HART, P. E. *Pattern Classification and Scene Analysis*. New York: Wiley-Interscience Publication, 1973. Citado na página 36.

ELIZABETH, D.; NEHEMIAH, H.; RAJ, C. R.; KANNAN, A. Computer-aided diagnosis of lung cancer based on analysis of the significant slice of chest computed tomography image. *IET Image Processing*, p. 697–705, 2012. Citado 2 vezes nas páginas 15 e 42.

FARAG, A.; ALI, A.; GRAHAM, J.; FARAG, A.; ELSHAZLY, S.; FALK, R. Evaluation of geometric feature descriptors for detection and classification of lung nodules in low dose ct scans of the chest. In: IEEE. *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. [S.l.], 2011. p. 169–172. Citado 2 vezes nas páginas 15 e 42.

FARAG, A. A.; ALI, A.; ELSHAZLY, S.; FARAG, A. A. Feature fusion for lung nodule classification. *International journal of computer assisted radiology and surgery*, Springer, v. 12, n. 10, p. 1809–1818, 2017. Citado 3 vezes nas páginas 17, 42 e 43.

FILHO, A. O. de C.; SAMPAIO, W. B. de; SILVA, A. C.; PAIVA, A. C. de; NUNES, R. A.; GATTASS, M. Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial Intelligence in Medicine*, v. 60, n. 3, p. 165–177, 2013. ISSN 0933-3657. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0933365713001541>. Citado 3 vezes nas páginas 16, 42 e 43.

FILHO, A. O. de C.; SAMPAIO, W. B. de; SILVA, A. C.; PAIVA, A. C. de; NUNES, R. A.; GATTASS, M. Automatic detection of solitary lung nodules using quality threshold clustering, genetic algorithm and diversity index. *Artificial Intelligence in Medicine*, v. 60, n. 3, p. 165–177, 2014. Disponível em: <http://dx.doi.org/10.1016/j.artmed.2013.11.002>. Citado 3 vezes nas páginas 36, 43 e 44.

FILHO, A. O. de C.; SILVA, A. C.; PAIVA, A. C. de; NUNES, R. A.; GATTASS, M. Lung-nodule classification based on computed tomography using taxonomic diversity indexes and an svm. *Journal of Signal Processing Systems*, p. 1–18, 2016. ISSN 1939-8115. Disponível em: <http://dx.doi.org/10.1007/s11265-016-1134-5>. Citado na página 43.

FISHMAN, A. P.; RONDINONE, S.; GUIOVANNIELLO, O. et al. *Fishman manual de enfermedades pulmonares*. [S.l.: s.n.], 2004. Citado na página 19.

FUJIMOTO, J.; WISTUBA, I. I. Current concepts on the molecular pathology of non-small cell lung carcinoma. *Seminars in Diagnostic Pathology*, v. 31, n. 4, p. 306 – 313, 2014. ISSN 0740-2570. Lung Carcinoma: Beyond The {WHO} Classification. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0740257014000616>. Citado na página 13.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 2nd. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1992. ISBN 0201508036. Citado 2 vezes nas páginas 20 e 21.

GORENSTEIN, M. R. *Diversidade de espécies em comunidades arbóreas: aplicação de índices de distinção taxonômica em três formações florestais do Estado de São Paulo*. Tese (Doutorado) — Universidade de São Paulo, 2009. Citado na página 25.

GOULD, M.; MACLEAN, C.; KUSCHNER, W.; RYDZAK, C.; OWENS, D. Accuracy of positron emission tomography for diagnosis of pulmonary nodules and mass lesions: A meta-analysis. *JAMA*, v. 285, n. 7, p. 914–924, 2001. Disponível em: <+http://dx.doi.org/10.1001/jama.285.7.914>. Citado na página 13.

HANSELL, D. M.; BANKIER, A. A.; MACMAHON, H.; MCLOUD, T. C.; MULLER, N. L.; REMY, J. Fleischner society: Glossary of terms for thoracic imaging. *Radiology*, v. 246, n. 3, p. 697–722, 2008. PMID: 18195376. Disponível em: <http://dx.doi.org/10.1148/radiol.2462070712>. Citado na página 13.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, Ieee, n. 6, p. 610–621, 1973. Citado na página 21.

HARDIE, R. C.; ROGERS, S. K.; WILSON, T.; ROGERS, A. Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. *Medical Image Analysis*, v. 12, n. 3, p. 240 – 258, 2008. ISSN 1361-8415. Disponível em: <http://www.sciencedirect.com/science/article/pii/S136184150700103X>. Citado na página 43.

INCA, I. N. d. C. *O que é o câncer?* 2019. Acessado em http://www1.inca.gov.br/conteudo. Citado na página 13.

JING, Z.; BIN, L.; LIANFANG, T. Lung nodule classification combining rule-based and svm. In: *Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on*. [S.l.: s.n.], 2010. p. 1033–1036. Citado na página 43.

KREWER, H.; GEIGER, B.; AL. et. Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography. *IEEE International Conference on Systems, Man, and Cybernetics*, p. 3887–3891, 2013. Citado 2 vezes nas páginas 16 e 42.

KUMAR, S. A.; RAMESH, D. J.; AL. et. Robust and automated lung nodule diagnosis from ct images based on fuzzy systems. *Process Automation, Control and Computing (PACC), 2011 International Conference on*, p. 1–6, 2011. Citado 3 vezes nas páginas 14, 42 e 43.

LACERDA, E. G. de; CARVALHO, A. de. Introdução aos algoritmos genéticos. *Sistemas inteligentes: aplicaçoes a recursos hidricos e ciências ambientais*, v. 1, p. 99–148, 1999. Citado na página 26.

LEDERLIN, M.; REVEL, M.-P.; KHALIL, A.; FERRETTI, G.; MILLERON, B.; LAURENT, F. Management strategy of pulmonary nodule in 2013. *Diagnostic and Interventional Imaging*, v. 94, n. 11, p. 1081 – 1094, 2013. ISSN 2211-5684. Disponível em:

<http://www.sciencedirect.com/science/article/pii/S2211568413001964>. Citado na página 13.

LEEF, J. L.; KLEIN, J. S. The solitary pulmonary nodule. *Radiologic Clinics of North America*, Elsevier, v. 40, n. 1, p. 123–143, 2002. Citado na página 13.

LI, S.; XU, P.; LI, B.; CHEN, L.; ZHOU, Z.; HAO, H.; DUAN, Y.; FOLKERT, M.; MA, J.; HUANG, S. et al. Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features. *Physics in Medicine & Biology*, IOP Publishing, v. 64, n. 17, p. 175012, 2019. Citado 3 vezes nas páginas 18, 42 e 43.

MAGURRAN, A. E. *Measuring biological diversity*. [S.l.]: John Wiley & Sons, 2013. Citado 2 vezes nas páginas 21 e 23.

MASOOD, A.; SHENG, B.; LI, P.; HOU, X.; WEI, X.; QIN, J.; FENG, D. Computer-assisted decision support system in pulmonary cancer detection and stage classification on ct images. *Journal of biomedical informatics*, Elsevier, v. 79, p. 117–128, 2018. Citado 3 vezes nas páginas 17, 42 e 43.

MELO, A. S. What do we win 'confounding' species richness and evenness in a diversity index? *Biota Neotrop.*, v. 8, 2008. Disponível em: <http://www.biotaneotropica.org.br/v8n3/en/abstract?point-of-view+bn00108032008.> Citado na página 41.

NASCIMENTO, L. B. Lung nodules classification in ct images using shannon and simpson diversity indices and svm. *Machine Learning and Data Mining in Pattern Recognition*, v. 7376, p. 454–466, 2012. Citado 7 vezes nas páginas 15, 18, 19, 25, 26, 30 e 42.

NIBALI, A.; HE, Z.; WOLLERSHEIM, D. Pulmonary nodule classification with deep residual networks. *International Journal of Computer Assisted Radiology and Surgery*, v. 12, n. 10, p. 1799–1808, Oct 2017. ISSN 1861-6429. Disponível em: <https://doi.org/10.1007/s11548-017-1605-6>. Citado 3 vezes nas páginas 17, 42 e 43.

NUNES, A. P.; SILVA, A. C.; PAIVA, A. C. de. Machine learning and data mining in pattern recognition: 6th international conference, mldm 2009, leipzig, germany, july 23-25, 2009. proceedings. In: _____. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. cap. Detection of Masses in Mammographic Images Using Simpson's Diversity Index in Circular Regions and SVM, p. 540–553. ISBN 978-3-642-03070-3. Disponível em: <http://dx.doi.org/10.1007/978-3-642-03070-3_41>. Citado na página 43.

OLIVEIRA, F. S. S. d. et al. Classificação de tecidos da mama em massa e não-massa usando índice de diversidade taxonômico e máquina de vetores de suporte. Universidade Federal do Maranhão, 2013. Citado na página 24.

OLIVEIRA, F. Soares Sérvulo de; FILHO, A. Oseas de C.; SILVA, A. C.; PAIVA, A. Cardoso de; GATTASS, M. Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and SVM. *Computers in Biology and Medicine*, v. 57, p. 42–53, fev. 2015. ISSN 00104825. Disponível em: <http://dx.doi.org/10.1016/j.compbiomed.2014.11.016>. Citado 5 vezes nas páginas 16, 31, 32, 43 e 44.

OROZCO, H. M.; VILLEGAS, O. O. V.; AL. et. Lung nodule classification in ct thorax images using support vector machines. p. 277–283, 2013. Citado 2 vezes nas páginas 16 e 42.

OROZCO, H. M.; VILLEGAS, O. O. V.; MAYNEZ, L. O.; SÁNCHEZ, V. G. C.; DOMINGUEZ, H. D. J. O. Lung nodule classification in frequency domain using support vector machines. In: IEEE. *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*. [S.l.], 2012. p. 870–875. Citado 2 vezes nas páginas 15 e 42.

PARVEEN, S. S.; KAVITHA, C. Classification of lung cancer nodules using svm kernels. *International Journal of Computer Applications*, v. 95, p. 25–28, 2014. Citado na página 42.

RICOTTA, C. A parametric diversity measure combining the relative abundances and taxonomic distinctiveness of species. *Diversity and Distributions*, Wiley Online Library, v. 10, n. 2, p. 143–146, 2004. Citado na página 23.

ROCHA, S. V. d.; JUNIOR, G. B.; SILVA, A. A. C. A.; PAIVA, A. C. d. Texture analysis of masses in digitized mammograms using Gleason and Menhinick diversity indexes. *Revista Brasileira de Engenharia Biomédica*, scielo, v. 30, p. 27 – 34, 03 2014. ISSN 1517-3151. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1517-31512014000100006&nrm=iso>. Citado na página 43.

SA, J.; DS, R.; JD., F. Content-based versus semantic-based retrieval: an LIDC case study. *SPIE Medical Imaging*, 2009. Citado na página 31.

SAMPAIO, W. B. de; SILVA, A. C.; PAIVA, A. C. de; GATTASS, M. Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, LBP and SVM. *Expert Syst. Appl.*, v. 42, n. 22, p. 8911–8928, 2015. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2015.07.046>. Citado na página 44.

SANTOS, V. K. d. Uma generalização da distribuição do índice de diversidade generalizada por good com aplicação em ciências agrárias. *Vanessa Kelly dos Santos.–2009*, v. 57, 2009. Citado na página 22.

SILVA, I. A. D.; BATALHA, M. A. Taxonomic distinctness and diversity of a hyperseasonal savanna in central brazil. *Diversity and distributions*, Wiley Online Library, v. 12, n. 6, p. 725–730, 2006. Citado na página 23.

SOUSA, U. S. et al. Classificação de massas na mama a partir de imagens mamográficas usando índice de diversidade de shannon-wiener. Universidade Federal do Maranhão, 2011. Citado na página 22.

SRICHAI, M. B.; NAIDICH, D. P.; AL. et. *Computed tomography and magnetic resonance of the thorax*. [S.l.]: Lippincott Williams & Wilkins, 2007. Citado na página 13.

VANDAMME, P.; POT, B.; GILLIS, M.; VOS, P. D.; KERSTERS, K.; SWINGS, J. Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiol. Mol. Biol. Rev.*, Am Soc Microbiol, v. 60, n. 2, p. 407–438, 1996. Citado na página 23.

VAPNIK, V.; VAPNIK, V. *Statistical learning theory*. [S.l.]: Wiley New York, 1998. Citado na página 25.

VIANA, G. V. R.; CEARÁ, F. Técnicas para construção de árvores filogenéticas. *Fortaleza: UFCE*, 2007. Citado na página 24.

VITT, L. J.; PIANKA, E. R. *Lizard ecology: historical and experimental perspectives*. [S.l.]: Princeton University Press, 2014. v. 290. Citado na página 22.

WEBB, C. O. Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *The American Naturalist*, v. 156, p. 145–155, 2000. Citado na página 25.

ZHANG, G.; YANG, Z.; GONG, L.; JIANG, S.; WANG, L. Classification of benign and malignant lung nodules from ct images based on hybrid features. *Physics in Medicine & Biology*, IOP Publishing, 2019. Citado 2 vezes nas páginas 18 e 42.

ZHAO, X.; LIU, L.; QI, S.; TENG, Y.; LI, J.; QIAN, W. Agile convolutional neural network for pulmonary nodule classification using ct images. *International journal of computer assisted radiology and surgery*, Springer, v. 13, n. 4, p. 585–595, 2018. Citado 3 vezes nas páginas 17, 42 e 43.

ZINOVEV, D.; FEIGENBAUM, J.; FURST, J.; RAICU, D. Probabilistic lung nodule classification with belief decision trees. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC*. [S.l.: s.n.], 2011. p. 4493–4498. ISSN 1557-170X. Citado 2 vezes nas páginas 14 e 42.